PROYECTO FINAL CARRERA DE INGENIERÍA EN TELECOMUNICACIONES

ESTUDIO DE TÉCNICAS DE MACHINE LEARNING APLICADAS A LA CLASIFICACIÓN DE CULTIVO EN IMÁGENES SATELITALES

Pablo Andrés Aguirre
Estudiante

Dr. Jorge Osmar Lugo
Director

Dr. Pablo Andrés Weder
Co-director

Miembros del Jurado
Dr.Germán Mato (Instituto Balseiro)
Dr. Damián Hernández (Instituto Balseiro)

9 de Diciembre de 2019

Área de Análisis y Sistemas Complejos - INVAP S.E.

Instituto Balseiro
Universidad Nacional de Cuyo
Comisión Nacional de Energía Atómica
Argentina

A mi familia
A mis amigos
A mi Rosita
Gracias por acompañarme
y apoyarme durante este
trayecto de mi vida.

Índice de símbolos

PCA	Principal Component Analysis
SAR	Synthetic Aperture Radar.
SVC	Support Vector Classifier.
SVM	Support Vector Machine.
MLP	Multi-Layer Perceptron.
OBIA	Object-Based Image Analysis.
S1	Sentinel 1
S2	Sentinel 2
NDVI	Normalized Difference Vegetation Index
RVI	Radar Vegetation Index
ESA	European Space Agency

Índice de contenidos

Índice de figuras

Índice de tablas

Resumen

El monitoreo de cultivo cumple un rol importante en la agricultura. Durante los últimos años se exploró el uso de algoritmos de aprendizaje supervisado (*machine learning*) en conjunto con imágenes satelitales de los cultivos a monitorear como una herramienta importante en la automatización de este proceso, pudiendo monitorear grandes extensiones de cultivo de manera rápida y eficiente.

En este trabajo se estudiaron los principios físicos del sensado remoto (con un enfoque en sensado satelital), los tipos de imágenes que se pueden recolectar de un terreno y las características específicas de cada uno. Se investigó cómo acceder a dichas imágenes, que son de dominio público. También se estudiaron distintos algoritmos de machine learning y sus bases matemáticas, junto con posibles métricas de rendimiento y sus características. Luego se aplicaron estos algoritmos a la clasificación de cultivos utilizando imágenes ópticas y de radar. Para esto se analizó cual sería la mejor manera de implementar estos algoritmos, utilizando finalmente el lenguaje Python con las implementaciones de los algoritmos encontradas en la librería Scikit-learn.

El objetivo del trabajo fue realizar una comparación del rendimiento de los algoritmos más utilizados en la clasificación de cultivo bajo distintas condiciones. Los métodos de clasificación principalmente estudiados fueron las técnicas de redes neuronales, máquinas de soporte vectorial, árboles de decisión y bosques aleatorios. Se trabajó buscando la mejor combinación de parámetros de cada clasificador mediante barridos aplicando cross-validation de 5 pliegues. Entre las condiciones estudiadas las más importantes fueron la clasificación considerando y no considerando al tiempo como una variable más, todo esto utilizando datos de distinta naturaleza (ópticos, radar o ambos).

En la clasificación con datos en los que no se tiene en cuenta la evolución temporal del cultivo el mejor clasificador obtenido utilizando datos ópticos fue una máquina de soporte vectorial con un *accuracy* del 91 % y un *kappa* de 0.87. El mejor clasificador entrenado con datos de radar también fue una máquina de soporte vectorial con *accuracy* de 73 % y *kappa* de 0.60.

Respecto a la clasificación con datos en los que se tiene en cuenta la evolución temporal, el mejor clasificador entrenado con imágenes ópticas fue una máquina de soporte vectorial con *accuracy* de 93 % y *kappa* de 0.91. El mejor clasificador obtenido

utilizando datos de radar fue un bosque aleatorio con *kappa* 0.86 y un *accuracy* de 95 %.

Se realizó un análisis del efecto que tiene el tipo de datos utilizado en la clasificación multitemporal de cultivos con períodos de siembra y cosecha similares, llegando a la conclusión de que los datos de radar son más efectivos que los ópticos en estos casos. Sin embargo utilizar la combinación de datos de radar y ópticos fue lo que mejor rendimiento obtuvo.

Por otro lado, se estudió el efecto de entrenar los algoritmos con una base de datos que posee un número de muestras distinto para cada clase. Esto llevó a la conclusión de que los algoritmos priorizan la correcta clasificación de las clases que presentan más muestras de entrenamiento. Este estudio indica que es muy importante que los datos de entrenamiento sean una muestra representativa del contexto en el que se van a utilizar los algoritmos.

Palabras clave: MACHINE LEARNING, CLASIFICACIÓN DE CULTIVO, SENTINEL 1, SENTINEL 2, CLASIFICACIÓN MULTITEMPORAL, CLASIFICACIÓN DE UNA SOLA FECHA

Abstract

Crop monitoring has an important role in agriculture. During the last years the use of machine learning algorithms with satellite imagery altogether was explored as an important tool in automation of this process because of its ability to monitor big extensions of field in a fast and efficient way.

In this work the physical principles of remote sensing were studied (with focus on satellite sensing) along with the possible types of imagery you can gather from agricultural terrain and specific characteristics of every one of this types. How to acquire this public domain images was investigated. Different machine learning algorithms were also studied along with their mathematical basis and possible performance metrics. Then this algorithms were applied to crop classification using optical and radar imagery. For this purpose the best way to implement said algorithms was analyzed. Finally coding language Python with the algorithms implementations found in Scikit-learn library were used.

The objective of this work was to make a comparation of the most used algorithms performances on crop classification under different conditions. The mainly studied methods were support vector machines, neural networks, decision trees and random forests. Research was made looking for the best combination of parameters for each classifier using 5-fold cross-validation. The most important conditions studied were multitemporal and single date classification, using data of different nature (optical, radar or both).

On single-date classification, the best classifier obtained using optical data was a support vector machine with accuracy of 91% and a kappa coefficient of 0.87. The best classifier trained with radar data was also a support vector machine with obtained accuracy of 73% and a 0.60 kappa coefficient.

Regarding multitemporal classification, the best classifier trained with optical data was a support vector machine with obtained accuracy of 93% and a 0.91 kappa coefficient. The best classifier obtained using radar data was a random forest with obtained kappa of 0.86 and accuracy of 95%.

An analysis of the effect of the type of data used on multitemporal classification of crops with similar planting and harvest periods was made. This study led to the conclusion that radar data is more effective than optical in this cases. However, using

a combination of both radar and optical data gave the best results.

On other side, the effect of trainning algorithms with a database that has a different number of samples for every class was studied. This led to the conclusion that the algorithms priorize the correct classification of the classes that have the most trainning samples. This study indicate that it is very important that the trainning data consist in a representative sample of the context the algorithms are going to be utilized.

Keywords: MACHINE LEARNING, CROP CLASSIFICATION, SENTINEL 1, SENTINEL 2, MULTITEMPORAL CLASSIFICATION, SINGLE-DATE CLASSIFICATION

Capítulo 1

Introduccion

1.1. Motivación y objetivo del trabajo

La administración estratégica de cultivos tiene repercusiones relevantes en el campo de la agricultura, por ejemplo, evitando el uso excesivo y agotamiento de fuentes de agua, disminuyendo la tasa de crecimiento de emisiones de gases contribuyentes al efecto invernadero y adoptando prácticas de conservación del suelo para garantizar demandas futuras de agricultura [1]. Para facilitar el proceso de toma de decisiones sobre la administración de tierra o la implementación de acciones agrarias, es crucial poder identificar y monitorear la distribución de cultivos. Los métodos manuales de monitoreo consisten en el sensado presencial de los campos, los que precisan de gente que se translade hasta los campos a monitorear. Estos métodos suelen requerir de altos costos de producción además de un ineficiente uso del tiempo [2]. El sensado remoto provee una solución a estos problemas poniendo en la mesa una opción más rápida y menos costosa, con posibilidad de un seguimiento regular a través de imágenes satelitales multiespectrales. Estas imágenes logran obtener la **firma espectral** de las plantas analizadas, es decir, el comportamiento de su reflectancia a lo largo de distintas longitudes de onda del espectro electromagnético. La facilidad de conseguir estas firmas espectrales puede ser combinada con técnicas de Machine Learning para lograr una clasificación de las imágenes obtenidas a nivel píxel de manera automática y eficiente.

Durante varios años se estudió el tema de la clasificación de cultivos a través del sensado remoto utilizando distintas fuentes de imágenes satelitales tales como Landsat, QuickBird o ASTER [3, 4]. Dentro de este tema, se observó que la información correspondiente al calendario de cultivo, patrones de cultivo, técnicas de administración y tamaño de parcelas debe ser incorporada a los algoritmos clasificadores para obtener mejores resultados [5]. Sin embargo, muchas de estas características escapan al alcance de la clasificación a nivel píxel de las imágenes, lo que limita su aplicación en la discriminación de cultivo.

La clasificación por lote de cultivo puede sobrepasar estos problemas uniendo píxeles adjacentes dentro de cada lote dentro de objetos homogéneos creados a través de un proceso de segmentación y después clasificando a nivel de objeto. Este enfoque, conocido como *Object-based image analysis* (OBIA), puede incorporar las características mencionadas anteriormente para así aumentar la efectividad de los algoritmos [6]. Otra fuente de error de los algoritmos es la similitud de la firma espectral entre distintos cultivos con patrones de desarrollo y calendarios de crecimiento comunes, un ejemplo de esto se discute en el Capítulo 5, donde se utilizó la información de la evolución temporal de los cultivos para clasificar. Para contrarrestar esto en los últimos años se investigó la incorporación de datos de radar, que iluminan el cultivo en otras longitudes de onda capaces de detectar características morfológicas de las plantas [7].

El objetivo de este trabajo fue estudiar y poner en práctica distintos algoritmos de Machine Learning para clasificar cultivos de soja, maíz y girasol utilizando imágenes multiespectrales del satélite Sentinel-2. Los algoritmos de clasificación utilizados se presentan en el Capítulo 2. Se compararon resultados obtenidos con cada algoritmo bajo distintas condiciones:

- Clasificación a nivel píxel con datos ópticos (Capítulo 3).
- Clasificación a nivel objeto utilizando datos ópticos sin tener en cuenta la variable temporal (Capítulo 4).
- Clasificación a nivel objeto utilizando datos ópticos teniendo en cuenta la evolución temporal de los cultivos (Capítulo 5).
- Últimas dos condiciones utilizando datos de radar SAR y su combinación con datos ópticos (Capítulo 6).

En cada caso se buscaron los mejores rendimientos de los algoritmos ajustando parámetros característicos de cada uno y se compararon los resultados con los últimos trabajos en el tema.

Entre los algoritmos utilizados se encuentran modelos de aprendizaje supervisado y no supervisado.

1.2. Sensado remoto

El término sensado remoto fue utilizado en primer lugar en principios de la década de 1960 para describir cualquier manera de observar la Tierra desde lejos, particularmente referido a fotografías aéreas [8]. En un contexto más amplio las actividades de sensado remoto incluyen un gran rango de aspectos, desde la base física para obtener

información a la distancia, a las plataformas de operación llevando el sistema de sensado, a la adquisición, almacenamiento e interpretación de datos. En este trabajo sólo se manipularan datos obtenidos a través de sensores satelitales.

Una observación remota requiere de algún tipo de interacción de energía entre el sensor y el objetivo. La señal detectada por el sensor puede ser energía solar reflejada por la superficie de la tierra o energía emitida propiamente por la superficie. Además de los receptores pasivos, existen sensores capaces de producir sus propios pulsos de energía, por lo que pueden observar la superficie terrestre sin importar las condiciones solares. En este trabajo se utilizaron datos captados por sensores pasivos (imágenes ópticas) y activos (imágenes SAR).

El sensado remoto también incluye el análisis e interpretación de los datos e imágenes adquiridas. Este es uno de los aspectos más importantes del sensado remoto dado que posibilita proveer información relevante en base a lo sensado.

1.2.1. Principios físicos

El sensado remoto tiene una fuerte base física, dado que implica recolectar señales electromagnéticas provenientes de objetos con diferentes propiedades físicas y químicas.

Tipos de sensado

Se pueden distinguir tres formas de sensar información de un objeto: por reflexión, por emisión, o por emision-reflexión combinadas. La primera es la más común porque utiliza la luz solar, la principal fuente de energía en la Tierra. El Sol ilumina la superficie y ésta refleja una porción de esta energía de vuelta al espacio dependiendo del tipo y composición de la cobertura presente en ella. Luego la energía reflejada es detectada por el sensor del satélite, quien graba y transmite esta señal a una estación receptora.

Las observaciones remotas también pueden estar basadas en la energía emitida desde la superficie de la Tierra, donde el sensor detecta la energía emitida desde la propia superficie, un ejemplo de esto serían las imágenes infrarrojas. Como no depende directamente del Sol, este tipo de observación puede realizarse durante el día y la noche.

Finalmente también se pueden realizar observaciones utilizando sensores activos, llamados así porque poseen su propia fuente de energía. Estos sensores son capaces de irradiar energía sobre los objetivos y luego grabar la reflección obtenida para caracterizarlos.

Espectro electromagnético

Dado que las fuentes de radiación son muy diversas y por lo tanto las radiaciones electromagnéticas varían desde longitudes de onda cortas a largas, se tiende a clasificar-

las en ciertos grupos de longitudes de ondas organizados dentro del llamado **espectro electromagnético** (Figura 1.1).

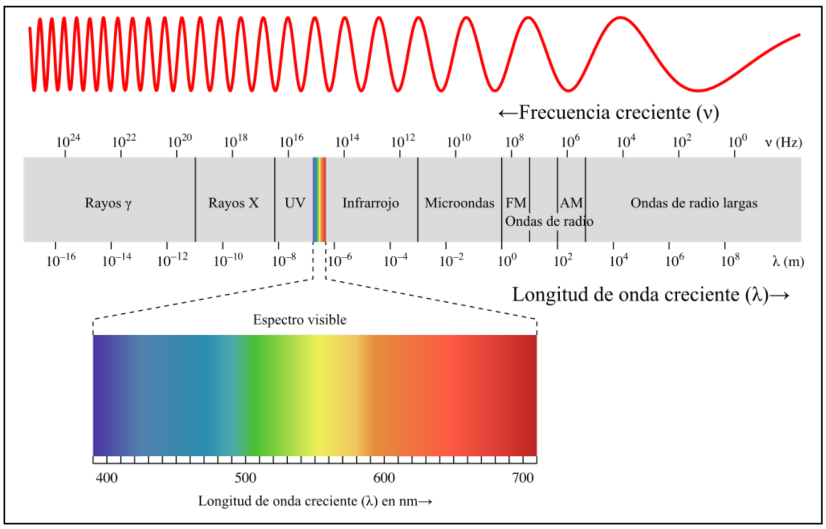

Figura 1.1: Espectro electromagnético con sus bandas espectrales más notables.

Las regiones espectrales comúnmente usadas en observacion remota son las siguientes:

- La región visible (0.4-0.7 μm). Cubre las longitudes de onda que nuestros ojos son capaces de detectar y a las cuales la energía del Sol es la mayor. Esta región puede ser subdividida en los tres colores primarios: azul (0.4-0.5 μm), verde (0.5-0.6 μm) y rojo (0.6-0.7 μm).

- La región infrarroja media (MIR, 1.2-8 μm). Esta región se encuentra entre las regiones del infrarrojo cercano (NIR) y del infrarrojo térmico (TIR). Desde 1.2 a 2.5 μm se encuentra la banda del infrarrojo de onda corta (SWIR), donde la influencia de la energía del Sol es aún muy relevante. Esta región provee las mejores estimaciones de la humedad del suelo y de la vegetación. De 3 a 8 μm la señal se convierte en una mezcla continua de energía solar reflejada y energía emitida por la superficie, volviéndose la componente emitida más relevante a medida que las longitudes de onda se agrandan. El intervalo de 3 a 5 μm es particularmente útil para detectar fuentes de alta temperatura, tales como volcanes o incendios forestales.

- La región infrarroja térmica (TIR de 8 a 14 μm). Esta región consiste en la energía emitida por la superficie terrestre, la cual es comúnmente usada para mapear temperaturas de superficie. La región térmica es usada para detectar evapotranspiración de vegetación, propiedades de hielo y nubes, efectos de calentamiento urbano y clasificación de rocas.

- La región de microondas (MW $>$ 1 cm). Esta región espectral es donde trabajan los sistemas de imagen radar. Su principal ventaja es la baja absorción atmosférica, lo que habilita a 'ver' a través de nubes. La radiación MW también puede penetrar follaje de bosques a varias profundidades y es útil en análisis de humedad del suelo y textura de superficie.

Reflexión y firma espectral

Se define radiancia (L) a la energía total que deja la superficie por unidad de tiempo y dentro de un ángulo sólido unitario (Ω), esta energía incluye tanto emisión como reflexión. Es una magnitud fundamental en sensado remoto debido a que describe exactamente lo que el sensor mide. La radiancia se expresa en watts por metro cuadrado por steradian ($Wm^{-2}sr^{-1}$). Reflectancia (ρ) es la relación entre la energía reflejada por una superficie y la energía incidente en la misma. La firma espectral de un objeto es el comportamiento de su reflectancia a lo largo de varias longitudes de onda del espectro electromagnético. En la Figura 1.2 se muestran firmas espectrales para distintos materiales [8].

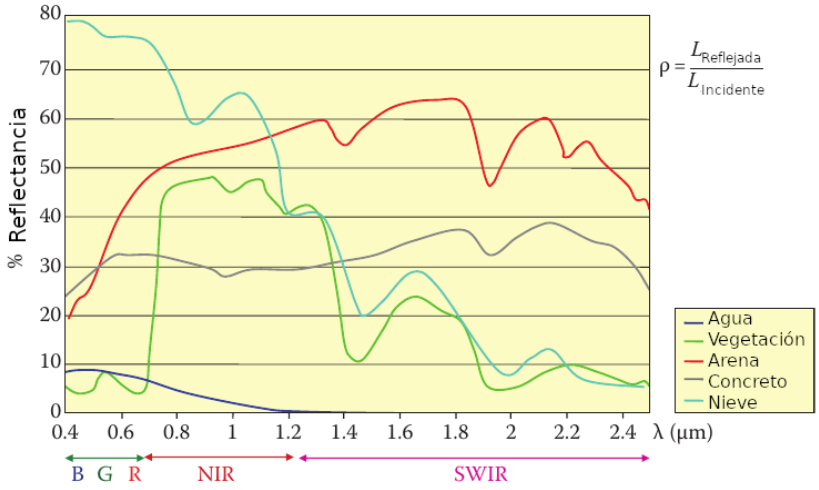

Figura 1.2: Firmas espectrales de reflectancias para materiales representativos de la superficie terrestre.

Las firmas espectrales son la base para poder distinguir objetos a partir de mediciones de sensado remoto en la región solar el espectro electromagnético. Sin embargo, estas firmas no son constantes para cada cobertura, ya que el flujo de radiancia detectado por los sensores no sólo depende de las propiedades intrínsecas del área observada, sino que también depende condiciones externas a la medición. Los principales factores que afectan las firmas espectrales son:

- Componentes atmosféricos, que afectan tanto la absorción como el scattering de radiacion incidente y reflajada.

- Variaciones medioambientales en la cubierta causantes de cambios en la composición física o química, tales como densidad, pigmentación, humedad o aspereza. Pueden ser causadas por la fenología del cultivo o vegetación, prácticas agriculturales, pasto, etc.

- Condiciones de iluminación solar, que varían con la latitud, fecha del año y hora de observación, además de posición del sensor.

- Pendiente del terreno.

1.3. Satélites y bandas utilizadas

En este trabajo se utilizaron datos correspondientes a mediciones realizadas de manera pasiva (imágenes ópticas) por el satélite Sentinel-2, y de manera activa por el radar perteneciente al Sentinel-1. Se utilizaron estos dos tipos de sensores debido a sus capacidades complementarias. En la Tabla 1.1 se muestran las características de estos sensores. En ella se puede apreciar que las imágenes opticas son mejores para detectar propiedades espectrales, mientras que las de radar son mejores para tener en cuenta características del follaje. También se pueden ver las distintas capacidades de detección ante condiciones climáticas o del terreno.

Características del cultivo		Datos ópticos	Datos SAR
Estructura del follaje	Tipo de hoja	-	+
	Altura	-	+
	Densidad	+	+
	Geometría	-	+
Propiedades espectrales	Pigmentos	+	-
	Color flores	+	-
	Reflexión NIR	+	-
Contenido de agua		+	+
Estimación de biomasa		+	+
Predicción de rendimiento		+	+
Etapas fenológicas		+	+
Características de la superficie del suelo		-	+
Fuentes de ruido o falta de datos			
Condiciones climáticas	Cubrimiento de nubes	-	+
	Efecto del viento en follaje	+	-
	Efectos atmosféricos	-	+
Humedad del suelo/Dry effect		+	+
Efecto de variaciones en el terreno		+	-

Tabla 1.1: Capacidades (+) y limitaciones (-) de datos ópticos y de SAR en el contexto de clasificación y monitoreo de cultivo.

1.3.1. Sentinel-1

Sentinel-1 es una constelación de satélites compuesta por los satélites Sentinel-1A y Sentinel-1B destinada al monitoreo terrestre y oceánica. Los satélites poseen un radar de apertura sintética (SAR) en la banda C operando a una frecuencia central de 5.405 GHz [9]. Estos satélites se encuentran en una misma órbita con altura orbital de 693 km, separados por 180 grados. Esta misión tiene un tiempo de revisita de 6 días.

Los radares tipo SAR buscan combinar la información obtenida en varios barridos de la antena para recrear un solo "barrido virtual". Al final el sistema radar proporciona el mismo rendimiento que daría si estuviese equipado con una antena más grande y directiva que la que tiene en realidad.

Las ondas de radar tienen una **polarización**. Diferentes materiales reflejan las ondas de radar con diferentes intensidades y normalmente la intensidad reflejada depende de la polarización de la onda incidente. Algunos materiales tambíen convierten una polarización en otra. Emitiendo una mezcla de polarizaciones y usando antenas receptoras con una polarización específica, varias imágenes diferentes pueden recolectarse de la misma serie de pulsos. Este radar soporta operación en polarización dual (HH+HV, VV+VH) implementada a través de una cadena de transmisión (conmutable a H o V) y dos cadenas paralelas de recepción para polarización H y V [9].

1.3.2. Sentinel-2

Sentinel-2 es otra misión de observación terrestre desarrollada por la ESA. A esta constelación la componen los satélites Sentinel-2A y Sentinel-2B, equipados con un instrumento multiespectral destinado a recolectar imágenes en distintas regiones del espectro. Estos satélites se encuentran en una órbita con altura orbital de 786 km e inclinación orbital de 98.62 grados, separados por 180 grados [10]. Esta misión tiene un tiempo de revisita de 5 días. En la Tabla 1.2 se detallan las bandas espectrales en las que se toman las imágenes.

1.3.3. Indices de vegetación

Las diferencias en la reflectancia obtenida de distintos cultivos pueden ser incrementadas mediante el uso de índices de vegetación. Estos índices consisten en combinaciones de valores de energía reflejada obtenidos en distintas bandas o polarizaciones que resultan en un valor indicador de alguna característica en especial. Un ejemplo es el NDVI, un índice que tiene relación con el tipo de vegetación y etapa de crecimiento de la planta. En la Tabla 1.3 se aprecian algunos índices importantes junto con sus definiciones.

En este trabajo se utilizaron los índices NDVI y RVI.

Resolución Espacial (m)	Número de banda	Longitud de onda central (nm)
10	2	496
	3	560
	4	664
	8	835
20	5	704
	6	740
	7	782
	8a	865
	11	1614
	12	2202
60	1	444
	9	945
	10	1373

Tabla 1.2: Bandas espectrales de la misión Sentinel-2.

Indices de vegetación	
NDVI	Utilizado para estimar cantidad calidad y desarrollo de la vegetación.
SAVI	Utilizado cuando la cubierta vegetal es baja.
GNDVI	Indice centrado en cantidad de clorofila detectada.
EVI	Incluye reducción de influencia atmosférica y es más sensible a cambios en el follaje.
RVI	Índice de vegetación de radar.

Tabla 1.3: Algunos índices de vegetación.

1.4. Segmentación

Uno de los pasos necesarios para poder trabajar a nivel objeto es la segmentación de las imágenes satelitales para delimitar los bordes de los campos de cultivo y calcular variables estadísticas para cada lote. En este trabajo el proceso de segmentación no se realizó dado que en INVAP se disponía de una base de datos resultante de haber aplicado este procesamiento a distintas imágenes en otro trabajo. La segmentación consiste en recortar los píxeles correspondientes a cada lote de cultivo gracias a que las posiciones geográficas de los límites de los lotes están disponibles en otra base de datos junto con el tipo de cultivo que se encuentra en ellos. Una vez aislados estos píxeles se realiza el cálculo de variables estadísticas de los valores encontrados, que luego se presentan en formas de features del lote.

1.5. Machine Learning

La clasificación de cultivos es una tarea que cae dentro de la categoría de *reconocimiento de patrones*. Esta clase de problemas son generalmente encarados mediante

el desarrollo de algoritmos de **machine learning** apropiados. Generalmente hablando, machine learning (o aprendizaje automático) involucra tareas para las cuales no existe un método directo conocido de computar una respuesta deseada a un conjunto de valores de entrada [11]. Dentro del machine learning existen dos tipos principales de métodos para resolver un problema, los métodos de aprendizaje supervisado y no supervisado.

1.5.1. Aprendizaje supervisado

La estrategia adoptada de los métodos de aprendizaje supervisado es que los algoritmos "aprendan" a partir de un conjunto de ejemplos representativos la relación que existe entre entrada y salida. Este conjunto de datos se denomina conjunto de entrenamiento. El aprendizaje consiste en que el algoritmo modifique variables específicas de cada modelo para conseguir emular los resultados del conjunto de entrenamiento teniendo como entradas los datos correspondientes. Luego de la etapa de entrenamiento el algoritmo se pone a prueba utilizando un conjunto distinto de datos etiquetados, llamado conjunto de testeo, y se mide su rendimiento en base a este conjunto.

Uno de los problemas del aprendizaje supervisado es el *overfit*, que ocurre cuando un algoritmo aprende las relaciones entre entrada y salida para el conjunto específico de entrenamiento, pero tiene un pobre desempeño a la hora de clasificar nuevos datos. En otras palabras, falla en generalizar el problema. Por esto es importante la etapa de entrenamiento, así como que el conjunto de entrenamiento sea representativo del problema en el que se va a utilizar el algoritmo.

1.5.2. Aprendizaje no supervisado

El aprendizaje supervisado involucra el uso de un set de datos de entrenamiento consistente en datos etiquetados y representativos de cada tipo de cultivo de interés. A diferencia de la clasificación supervisada, la clasificación no supervisada no requiere información de referencia para ser realizada. En lugar de eso, se intenta encontrar una estructura subyacante de clase auntomáticamente organizando los datos en grupos que comparten características similares. Normalmente sólo es necesario especificar de antemano el número K de clases presentes.

La clasificación no supervisada juega un papel importante cuando hay disponible poca información *a priori* de los datos. Uno de los objetivos principales de usar algoritmos de aprendizaje no supervisado para datos de sensado remoto multiespectral es obtener información útil para la selección de regiones de entrenamiento en una clasificación supervisada subsecuente.

1.5.3. Métricas

Uno de los puntos claves a la hora de trabajar en un problema de machine learning es elegir una métrica adecuada para evaluar el desempeño del algoritmo. En este trabajo se utilizaron tres métricas distintas: F1-score, Kappa, y la exactitud o accuracy.

Accuracy

El accuracy o exactitud es simplemente una medida de cuántos ejemplos se clasificaron correctamente del total. Se calcula como:

$$Accuracy = \frac{Ejemplos\ correctamente\ clasficados}{Total\ de\ ejemplos}$$

F1-Score

El puntaje F1 es una métrica que a su vez tiene en cuenta otras dos métricas: *precision* y *recall*. Precision es una medida de cuántos ejemplos clasificados en una clase son realmente pertenecientes a ella, y recall es una medida de cuántos ejemplos pertenecientes a una clase fueron clasificados correctamente. Para poder entender mejor en qué consisten estas métricas se presentan en la Tabla 1.4 definiciones de ejemplos bien y mal clasificados en un escenario de valores positivos y negativos.

		Predicción	
		Negativo	**Positivo**
Real	**Negativo**	Verdadero Negativo	Falso Positivo
	Positivo	Falso Negativo	Verdadero Positivo

Tabla 1.4: Comparación entre valores predichos y reales junto con su denominación.

En el caso en el que se trabajó, un Verdadero Positivo sería un ejemplo clasificado dentro de una clase a la que realmente pertenece. Teniendo en cuenta las definiciones de la Tabla 1.4 se definen:

$$Precision = \frac{Verdaderos\ Positivos}{Total\ de\ positivos\ predichos}$$

$$Recall = \frac{Verdaderos\ Positivos}{Total\ de\ positivos\ reales}$$

$$F1 - Score = 2\ \frac{Precision\ *\ Recall}{Precision\ +\ Recall}$$

Donde el F1-Score es la media armónica[1] entre precision y recall. Esta métrica se utiliza cuando se busca un balance entre precision y recall. Esta métrica es una mejor

[1]La media armónica de una cantidad finita de números es igual al recíproco, o inverso, de la media aritmética de los recíprocos de dichos valores.

manera de medir desempeño que el accuracy en casos en los que no hay una distribución pareja de ejemplos de entrenamiento.

Kappa

La métrica Kappa es una métrica que compara la exactitud observada con una exactitud esperada de un clasificador aleatorio. Por esto puede ser una elección menos engañosa que la exactitud como métrica (una exactitud observada del 80 % es menos impresionante con una exactitud esperada del 75 % que con una exactitud esperada del 50 %). Kappa se calcula como [12]:

$$\kappa = \frac{Exactitud\ observada - Exactitud\ esperada}{1 - Exactitud\ esperada}$$

1.5.4. K-fold Cross Validation

Como fue mencionado en 1.5.1 uno de los principales problemas a enfrentar trabajando con aprendizaje supervisado es la generalización del rendimiento del algoritmo ante distintos sets de testeo. *K-fold Cross Validation* (CV) provee una solución a este problema dividiendo los datos disponibles en subconjuntos y asegurándose de que cada subconjunto es usado como test set en algún momento. En el trabajo se utilizó CV con 5 subdivisiones o folds.

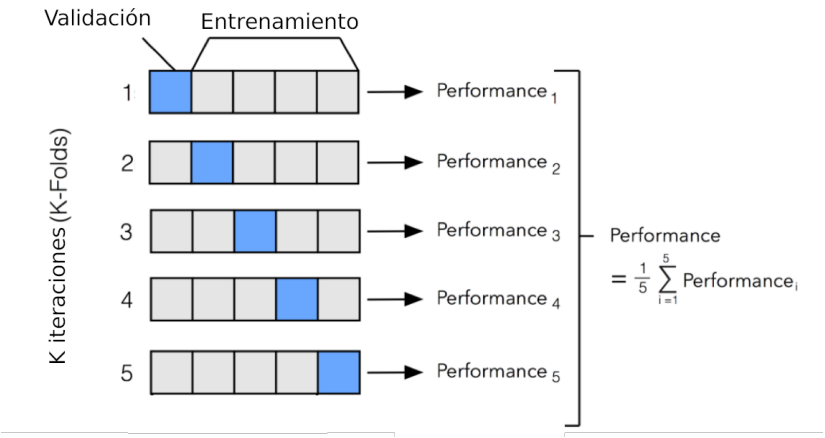

Figura 1.3: 5-Fold Cross-Validation.

La manera en la que trabaja este método es la siguiente: Se separa un test set final de los datos, luego al set de entrenamiento restante se lo divide en 5 subconjuntos y se entrena al algoritmo utilizando cada uno de estos subconjuntos como test set (Figura 1.3). Una vez realizado esto, se realiza la evaluación final del algoritmo utilizando el test set separado en el primer paso. El número de folds a utilizar en la CV se elige teniendo

en cuenta la cantidad de muestras que quedan en el set de testeo. Estas muestras tienen que conformar un subconjunto representativo del total de muestras disponible.

Esta metodología de trabajo es buena para evitar overfitting y también hace un buen uso de los datos disponibles, algo crucial cuando se dispone de pocos datos de entrenamiento. También se la utiliza para optimizar sobre hiperparámetros de los algoritmos. Es decir, uno elige el hiperparámetro que tiene mejor desempeño en los tests de CV, y luego se evalúa el desempeño en el test set final (el separado desde el comienzo).

Capítulo 2

Técnicas de clasificación usadas

Durante el trabajo se realizó el estudio y la comparación entre distintos métodos de clasificación. Estos métodos pueden ser separados en dos grupos: los **metódos de clasificación supervisada y no supervisada**. Los métodos de clasificación no supervisada se basan en la agrupación de datos en distintos grupos o *clusters* sin la necesidad de información de referencia para realizar esta agrupación. Los métodos de clasificación supervisada se caracterizan por requerir una base de datos previamente clasificada para realizar el entrenamiento del clasificador.

2.1. Métodos no supervisados

La clasificación no supervisada cumple un rol importante en los problemas en los que se dispone de poca información *a priori* sobre los datos. Normalmente a un algoritmo de clasificación no supervisada lo único que se le especifica de antemano es la cantidad K de clusters en los que se quiere agrupar los datos [11]. En los problemas de clasificación de cultivo este tipo de métodos no es muy utilizado debido a su bajo rendimiento, por lo que se implementará sólo un algoritmo de este tipo.

2.1.1. K-means

Para el problema de K-means se suministra un entero k, y un set de n puntos de datos $\mathbf{X}$. Se desea elegir k centros C para minimizar la función potencial [13],

$$\phi = \sum_{x \in X} \min_{c \in C} \|x - c\|^2 \tag{2.1}$$

A partir de estos centros, se agrupan puntos de datos de acuerdo a qué centro está mas cercano cada punto. El algoritmo de k-means es simple y rápido para la resolución del problema de agrupación. Sus pasos son:

1. Elegir arbitrariamente k centros iniciales $C = \{c_1, c_2, ..., c_k\}$.

2. Para cada $i \in \{1, ..., k\}$, elegir el cluster C_i al set de puntos en $\mathbf{X}$ que se encuentran más cercanos a c_i que a c_j para todo $j \neq i$.

3. Para cada $i \in \{1, ..., k\}$, elegir c_i como el centro de masa de todos los puntos en $C_i : c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$.

4. Repetir pasos 2 y 3 hasta que C no cambie más o los cambios sean menores a un margen especificado.

Dado suficiente tiempo el algoritmo siempre converge, sin embargo puede hacerlo en un mínimo local. Esto es altamente dependiente de la inicialización de los centroides. Como resultado la computación es normalmente realizada múltiples veces, con diferentes inicializaciones de los centroides, tomando como finales los centroides que minimicen 2.1.

2.2. Métodos supervisados

La clasificación supervisada se puede ver como un problema de modelado de distribuciones condicionales de probabilidad [11]. Basándose en datos representativos de K clases, se "aprenden" o aproximan las probabilidades *a posteriori* para la clase k condicional a la observación $\mathbf{g}$, $Pr(k|\mathbf{g}), k = 1, ..., K$. A este proceso se lo denomina normalmente como la *fase de entrenamiento* del proceso de clasificación. Luego de la fase de entrenamiento, se utilizan las probabilidades aproximadas para clasificar los datos que se ingresen al clasificador.

A la hora de investigar sobre los distintos tipos de métodos de este tipo, se encontró que los mayormente usados son unos cuantos, siendo estos:

- Máquinas de Soporte Vectorial.
- Regresión Lineal.
- Regresión Logística.
- Naive Bayes.
- Árboles de Decisión y Bosques Aleatorios.
- K-Vecinos Cercanos.
- Redes Neuronales (Multilayer Perceptron).

Sin embargo, los algoritmos utilizados para la clasificación de cultivo en los estudios encontrados sobre el tema [5–7, 14–21] son en general las redes neuronales, las máquinas de soporte vectorial, y los bosques aleatorios. Estos algoritmos se usan debido a su probada eficacia en la clasificación de cultivos. En este trabajo se utilizaran dichos algoritmos, los que se presentan a continuación junto con sus características.

2.2.1. Support Vector Machine

Las Máquinas de Soporte Vectorial (*Support Vector Machines* o SVMs) son un conjunto de métodos de aprendizaje usados para clasificación, regresión y detección de valores atípicos [22]. Una SVM construye un hiperplano en un espacio de gran dimensionalidad para separar este espacio en dos subespacios, uno para cada clase. El método busca optimizar este hiperplano para conseguir el mayor margen posible, es decir, la mayor distancia a los puntos de entrenamiento más cercanos. En la Figura 2.1 se muestra un ejemplo de hiperplano.

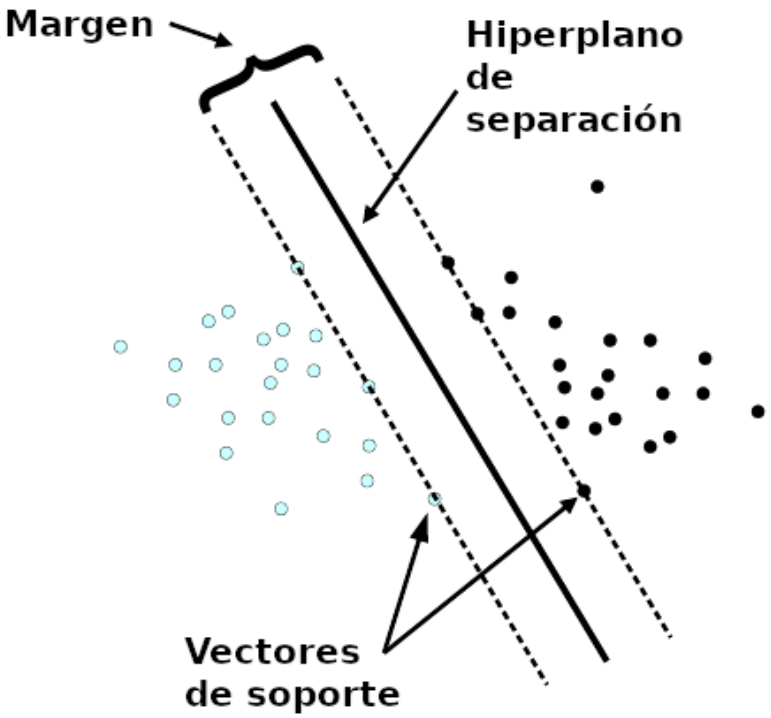

Figura 2.1: Hiperplano, margen y vectores de soporte de una clasificación por SVM.

Formulación matemática

Dados vectores de entrenamiento $\mathbf{x}_i \in \mathbb{R}^p, i = 1, ..., n$ pertenecientes a dos clases, y un vector $\mathbf{y} \in \{1, -1\}^n$ un clasificador de vectores de soporte (SVC) resuelve el

siguiente problema:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \tag{2.2}$$

Sujeto a:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i$$

$$\zeta_i \geq 0, i = 1, ..., n$$

Donde ζ_i es una variable que depende de si el dato x_i está bien clasificado o no, C regula el impacto de los ejemplos de entrenamiento en la maximización del margen de la función de decisión, ϕ es una transformación de los puntos de entrenamiento hacia un espacio de dimensionalidad mayor y w es el vector normal al hiperplano. A esta transformación se le llama *kernel*.

Los kernels utilizados en este trabajo fueron *Radial Basis Function* (RBF) y *Polynomial*.

Una propiedad importante de las SVMs es que la determinación de los parámetros del modelo corresponde a un problema de optimización de una función convexa. Dado esto, cualquier solución local es también un mínimo global de la función.

Entre las ventajas de las SVMs se encuentran su efectividad en espacios de alta dimensionalidad, el uso de puntos de entrenamiento en la función de decisión (vectores de soporte), y su versatilidad en función de los kernel que se pueden especificar para la función de decisión. Sus desventajas incluyen que es necesaria una buena elección de kernel y de regularización si el número de features es mucho mayor al número de muestras para evitar overfitting, y que este modelo no provee directamente estimados de probabilidad, sino que deben ser calculados utilizando CV.

2.2.2. Redes Neuronales

Las redes neuronales son algoritmos ampliamente utilizados debido a su capacidad de resolver relaciones no lineales entre datos de entrada y salida. Un *Multi-layer Perceptron* (MLP) es un algoritmo de aprendizaje supervisado que aprende una función $f(.) : R^m \rightarrow R^o$ entrenando en un set de datos, donde m es el número de dimensiones de entrada y o el número de dimensiones de salida.

Un MLP consiste en capas de neuronas interconectadas entre sí. En la Figura 2.3 se muestra una red neuronal. Cada neurona realiza una combinación lineal de sus elementos de entrada, a esto le suma un valor distinto para cada neurona llamado bias y al resultado lo hace pasar por una función no lineal llamada **función de activación**. En este trabajo la función de activación utilizada es la ReLU, una función que vale 0 para valores de entrada negativos, y es igual al valor de entrada para valores mayores o iguales a 0. Una representación de neurona se muestra en la Figura 2.2

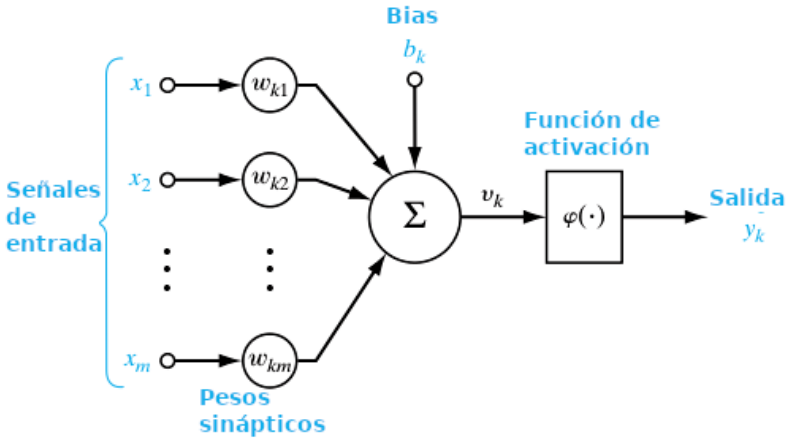

Figura 2.2: Representación de una neurona.

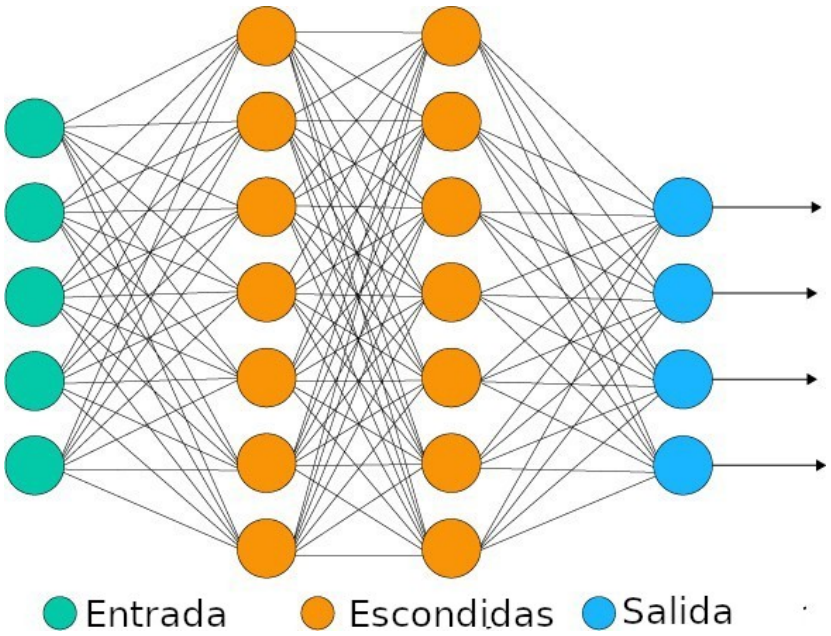

Figura 2.3: Red neuronal. Se distinguen las capas de entrada, salida, y las intermedias.

Formulación matemática

El modelo básico de neurona construye una combinación lineal de los valores de entrada y los hace pasar por una función de activación. La salida de la neurona j-ésima perteneciente a la capa l-ésima de la red, donde sus entradas son los valores $x_1, x_2, ..., x_n$ es:

$$z_j^{(l)} = g(\sum_{i=1}^{n} w_{ji}^{(l)} x_j + b_j^{(l)}) \tag{2.3}$$

Donde $g(.)$ es la función de activación de la neurona, $w_{ji}(l)$ es el peso correspondiente al valor i-ésimo de entrada a la neurona, y $b_j^{(l)}$ es el bias correspondiente a la neurona.

En una red neuronal las entradas a las neuronas de una capa son las salidas de las neuronas de la capa anterior. La capa de entrada consiste en los valores de entrada a

la red. En el caso de clasificación con más de dos clases, en la salida de la red se aplica la función *softmax*:

$$softmax(z_i) = \frac{\exp(z_i)}{\sum_l \exp(z_l)} \tag{2.4}$$

Esta función normaliza la salida para que a cada clase se le asigne un valor entre 0 y 1, tal que la suma de estos valores sea 1. Por esto se suele decir que la capa de softmax da resultados interpretables como probabilidades de que el dato de entrada pertenezca a cada clase.

El entrenamiento de la red se realiza actualizando los parámetros del modelo para minimizar la función de pérdida calculada a partir de los datos de entrenamiento. Estos parámetros son los pesos de las neuronas, los bias y parámetros de regularización. La suma de dichos parámetros puede resultar en un gran número de variables a ajustar, lo que puede llegar a ser una desventaja. La función de pérdida utilizada para clasificación es la *Cross Entropy Loss Function*:

$$Loss = \sum_{i}^{n} \sum_{k}^{K} -y_{true}^{(k)} \log(y_{predicted}^{(k)}) \tag{2.5}$$

Donde n es la cantidad de datos de entrenamiento, K es la cantidad de clases, $y_{predicted}$ es la salida de la red y y_{true} la esperada. A esta función normalmente se le agregan términos de regularización, cuya función principal es agregar algún parámetro que controle el overfitting. En este trabajo se utilizará regularización L2 controlada por el parámetro α [23]. Esta regularización consiste en un sesgo a la función 2.5 , mediante un término proporcional a los pesos de entrada en las neuronas. La función de pérdida utilizada con regularización L2 es:

$$Loss = \sum_{i}^{n} \sum_{k}^{K} -y_{true}^{(k)} log(y_{predict}^{(k)}) + \alpha \left\|\mathbf{W}\right\|_2^2 \tag{2.6}$$

Donde $\mathbf{W}$ es la matriz de pesos de las neuronas, y α el parámetro de regularización.

2.2.3. Decision Tree y Random Forest

Un *decision tree* (DT, árbol de decisión) es una clase de modelo de aprendizaje capaz de conseguir alta exactitud en distintas tareas y ser fácilmente interpretable. Este modelo toma decisiones de manera jerárquica, lo que lleva a una estructura de decisión entendible. El algoritmo trabaja separando los datos basandose en condiciones aplicadas a los valores de distintas features maximizando con cada separación la **ganancia de información** obtenida. Esta ganancia se mide a partir de la entropía o el índice Gini, mostrados en la Tabla 2.1. Un ejemplo de árbol de decisión se muestra en la Figura 2.4.

	Expresión matemática
Indice Gini	$1 - \sum_{i=1}^{C}(p_i)^2$
Entropía	$\sum_{i=1}^{C} -p_i \ log_2(p_i)$

Tabla 2.1: Medidas para calcular ganancia de información. p_i es la probabilidad de que un objeto sea clasificado en la clase i.

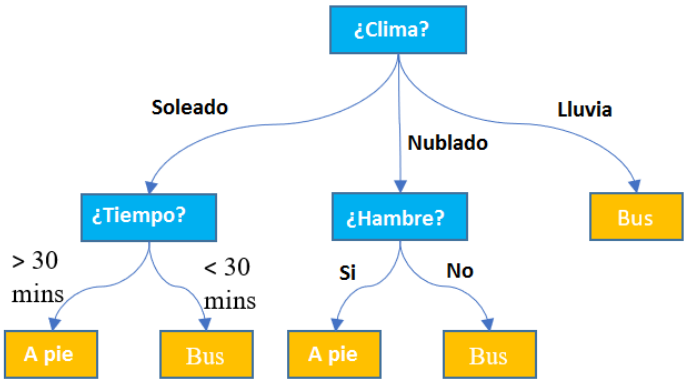

Figura 2.4: Decision tree con el problema de si viajar en bus o a pie.

Un clasificador *Random Forest* (RF) consiste en un grupo de árboles de decisión descorrelacionados donde las respuestas de los árboles se promedian para dar una respuesta final [24]. Este clasificador mostró buenos resultados en un número de aplicaciones de sensado remoto [25] [14].

Este algoritmo posee múltiples beneficios en relación a otros modelos de aprendizaje, entre los beneficios destacables se encuentran [15]:

- RF es capaz de correr en grandes conjuntos de datos.
- Es robusto al ruido y a valores atípicos.
- La complejidad computacional de RF es baja comparada con otros métodos.

Capítulo 3

Clasificación a nivel píxel

Como una forma de entrar en el tema de clasificación y elección de rutinas o librerías de software libre la primera tarea que se realizó en el proyecto fue la implementación de clasificadores a nivel píxel utilizando imágenes satelitales. Para esta tarea se utilizaron imágenes obtenidas de LandViewer, una interfaz web proporcionada por *Earth Observation System* (EOS).

Existen muchas librerías y lenguajes en los que se puede realizar este tipo de trabajos. Entre las herramientas más conocidas se encuentran Matlab, Python, Tensorflow, Java, R y C++. Se utilizó el lenguaje de programación Python por ser de libre uso con una gran comunidad de usuarios. Las librerías específicas de machine learning utilizadas se encuentran todas dentro del paquete scikit-learn.

3.1. Análisis de imágenes

El primer paso realizado fue una verificación de manera visual de la diferencia en la reflectividad de distintos elementos dentro de diferentes bandas. Para este acercamiento se utilizaron imágenes de la misión Landsat 8 debido a que la mayoría de sus bandas tienen la misma resolución espacial [26], lo que facilita su manipulación a nivel píxel. Se visualizaron distintas imágenes utilizando diferentes combinaciones de bandas como colores RGB. Un ejemplo de las imágenes utilizadas se muestra en la Figura 3.1, una imagen normal utilizando las bandas R,G y B de la costa de la provincia de Buenos Aires donde se pueden diferenciar zonas urbanas, de vegetación, y marítimas. Luego, en la Figura 3.2 se puede ver una imagen de la misma zona utilizando las bandas 7, 6 y 5 (bandas pertenecientes a las zonas NIR y SWIR del espectro) como rojo, verde y azul correspondientemente. Se puede apreciar que el agua absorbe la mayoría de la energía incidente dentro de este rango espectral, por lo que en la imagen esa parte se ve negra.

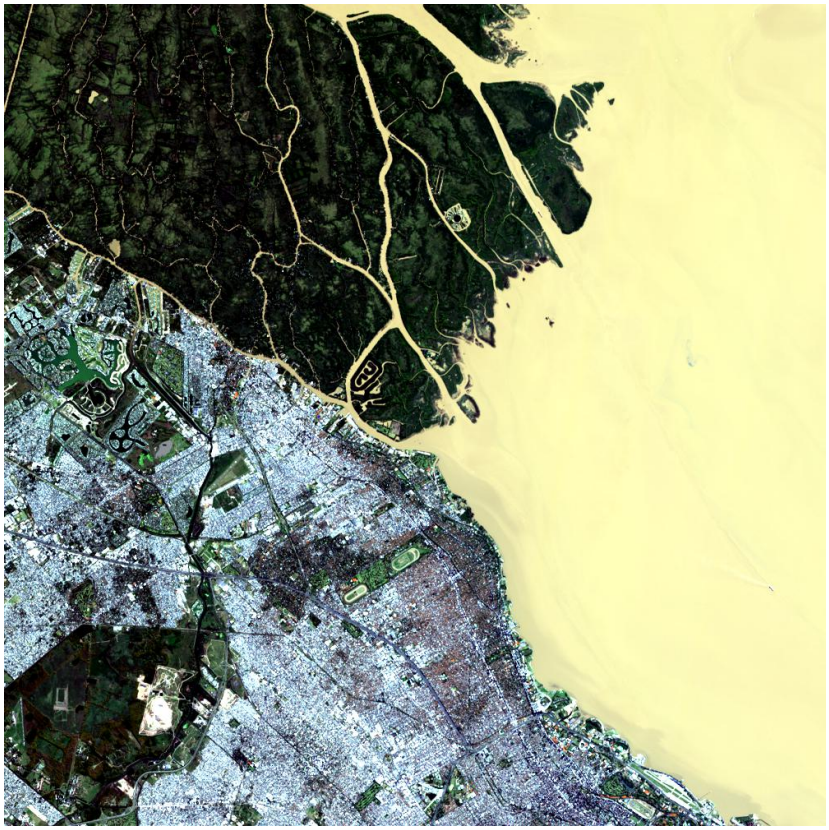

Figura 3.1: Imagen utilizada para comparar combinaciones de bandas y aspectos que éstas resaltan.

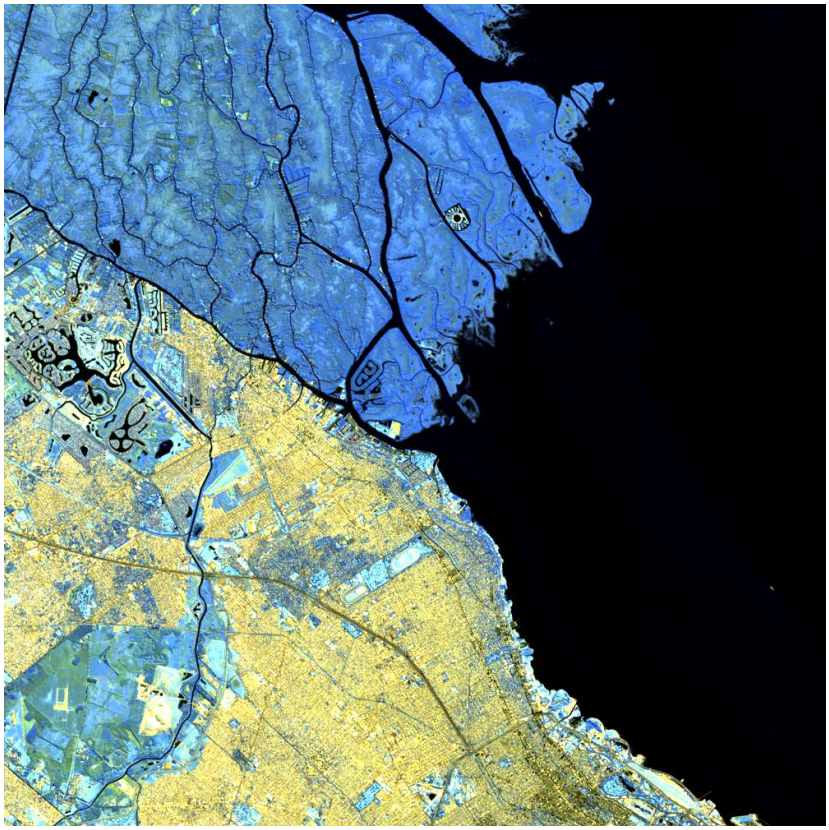

Figura 3.2: Imagen obtenida al utilizar las bandas 7, 6 y 5 como colores rojo, verde y azul.

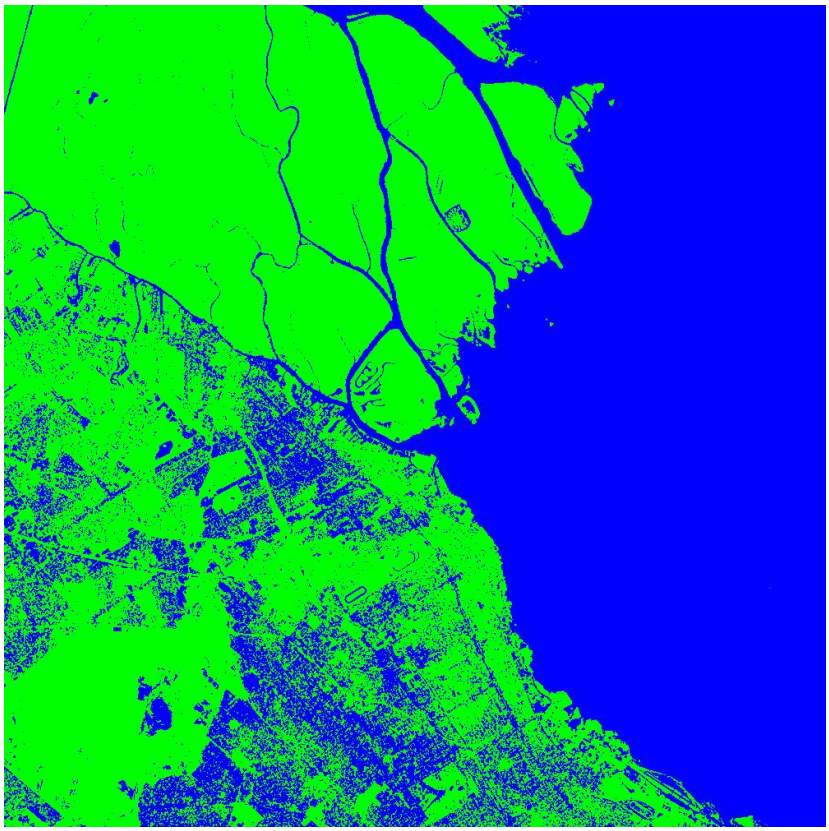

Figura 3.3: Grupos obtenidos al realizar la clasificación con K = 2.

3.2. K-Means

Luego de confirmar visualmente las diferentes características espectrales de los elementos presentes en la imagen de prueba se procedió a realizar una clasificación no supervisada de los píxeles utilizando todas las bandas disponibles con resolución de 30 metros. En primer lugar se buscó lograr la separación de lo que es agua de lo que no, así que se realizó una clasificación con K = 2 grupos. En la Figura 3.3 se muestran los resultados. En ella se ve que la agrupación se realizó de manera similar a lo que se esperaba, y que los píxeles agrupados en la categoría azul son mayormente los ubicados en zonas con agua.

Luego de realizar una primera agrupación de los píxeles se buscó realizar otra dividiendo la imagen en 3 grupos para ver si la nueva categoría agrupaba a la zona perteneciente a la ciudad. Los resultados se muestran en la Figura 3.4, donde se puede ver que agregando una tercera categoría los píxeles pertenecientes a esta están, en su mayoría, ubicados en la zona urbana.

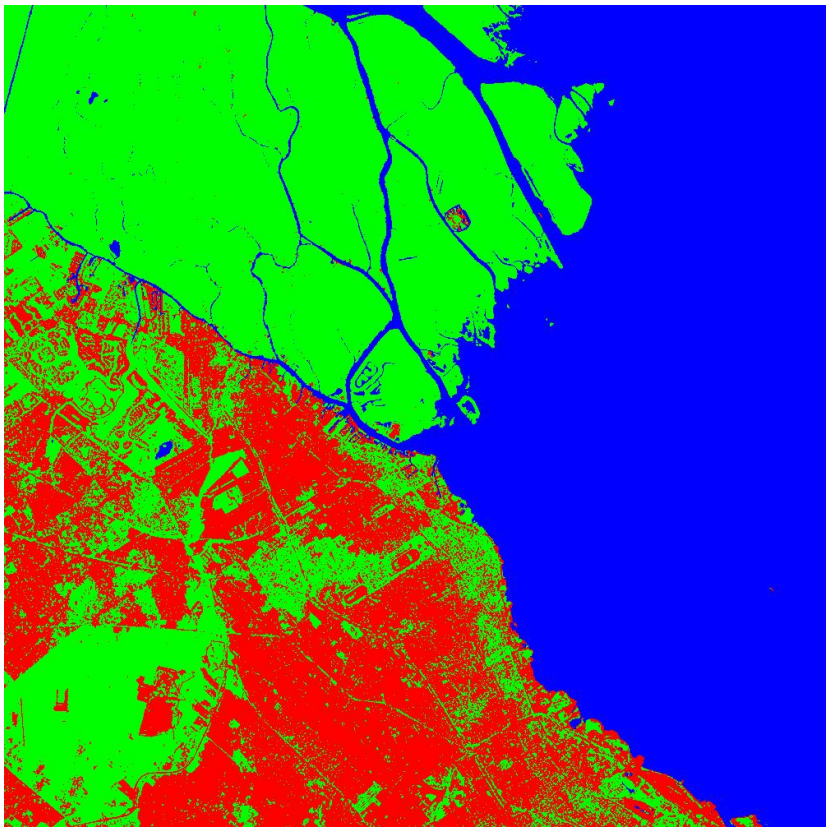

Figura 3.4: Grupos obtenidos al realizar la clasificación con K = 3.

3.3. Aprendizaje supervisado

Luego de ver el desempeño del algoritmo de K-Means en la imagen de prueba se decidió probar los algoritmos de aprendizaje supervisado. Para esto se recortaron tres imágenes de 150x200 píxeles pertenecientes a los grupos de clasificación deseados (vegetación, ciudad y agua) y se etiquetaron los píxeles para formar una pequeña base de datos. Con los datos etiquetados se procedió a entrenar los algoritmos de aprendizaje supervisado con los parámetros definidos por defecto y revisar su eficacia. Los resultados obtenidos se muestran en la Tabla 3.1.

Modelo	Accuracy
SVM	0.65
MLP	0.97
Decision Tree	0.98
Random Forest	0.98

Tabla 3.1: Resultados obtenidos con los algoritmos de aprendizaje supervisado trabajando a nivel píxel.

Se puede observar que los algoritmos tuvieron un buen rendimiento en general. Sin embargo, dada la resolución de 30 metros de las imágenes, por cada hectárea bajo análisis se necesitan alrededor de 11 píxeles con 8 bandas cada uno. Esto significa que

para poder entrenar los modelos de clasificación con datos pertenecientes a un espacio total de gran tamaño la cantidad de elementos a manipular se volvería considerablemente grande, y esto llevaría a que el entrenamiento requiera de mucho tiempo y poder de cómputo. La base de datos con la que se desea trabajar contiene información correspondiente a lotes de cultivo que abarcan en conjunto aproximadamente 2 millones de hectáreas. Realizando un cálculo rápido, se puede estimar que la base de datos necesaria para abarcar ese área necesitaría de alrededor de 22 millones de píxeles con 8 features por píxel.

Si se combina el problema del tamaño de la base de datos con el problema de que hay características del cultivo, como los patrones de arado, o la varianza entre píxeles de un mismo lote, que no se pueden apreciar a nivel píxel, se puede llegar a la conclusión de que clasificar los lotes de cultivo a nivel píxel no es una opción tan buena y sería conveniente analizar los lotes a nivel objeto, donde las características de píxeles adyacentes pertenecientes a una zona de interés se integran dentro de una muestra con variables estadísticas de la distribución de píxeles que la integran.

3.4. Conclusiones

Se realizó el acercamiento a Python y las librerías de scikit-learn para lograr la implementación de clasificadores supervisados y no supervisados a nivel píxel. Se consiguieron clasificadores con un buen rendimiento pero el problema que se afrontó en este capítulo no es el central del proyecto integrador, que es clasificar tipos de cultivo.

Se encontró una relación entre el tamaño del área de entrenamiento y el de la base de datos correspondiente. Se discutieron las ventajas de clasificar a nivel objeto respecto a clasificar a nivel píxel.

Capítulo 4

Datos ópticos: Clasificación sin tener en cuenta la variable temporal

Una vez que se realizó un primer acercamiento a los algoritmos realizando clasificación a nivel pixel, se decidió pasar a trabajar a nivel objeto para ya abordar el problema central del trabajo que es lograr la clasificación de distintos tipos de cultivo.

La clasificación a nivel objeto consiste en la agrupación de píxeles en segmentos por zona de interés, en este caso lotes de cultivo [5]. Esto posee numerosas ventajas frente a la clasificación a nivel pixel, como ser la información espectral adicional que poseen los segmentos (por ejemplo: valores medios por banda, medianas, minimos, máximos, varianza, etc.). Una ventaja aún mayor que la diversificación de los valores descriptivos de un objeto es la información espacial adicional que se obtiene [6]. Esta información espacial es de gran utilidad cuando se quiere utilizar información presente en la evolución temporal de las características observadas.

Los tipos de cultivo a clasificar fueron soja, maíz de primera cosecha, maíz de segunda cosecha (mismo maíz que el de primera, pero sembrado en otro momento del año) y girasol. Para el entrenamiento de los algoritmos se utilizó una base de datos brindada por INVAP con datos etiquetados correspondientes a distintos lotes de cutivo de Argentina. En esta parte del trabajo no se tuvo en cuenta la dimensión temporal de los datos.

4.1. Análisis de la base de datos

Los datos presentes en la base de datos fueron obtenidos gracias al sometimiento de imágenes multiespectrales capturadas por Sentinel-2 a un preprocesamiento donde se realizó la calibración de las imágenes, segmentación, numeración, clasificación y la obtención de variables estadísticas en cada banda utilizada. En la Tabla 4.1 se muestran los datos obtenidos para cada muestra de entrenamiento, y en la Tabla 4.2 se especifican

las valores que se calcularon para cada banda de frecuencia. La clasificación de la base de datos se realizó de manera manual mediante el sensado presencial de los cultivos.

En la base de datos se encuentran un total de 168000 muestras, pertenecientes a lotes de las provincias de Córdoba, Sana Fe, Buenos Aires, San Luis y Santiago del Estero. Las muestras son de clases distribuidas de la siguiente manera:

- Muestras de soja: 130849 (77.90 % del total)
- Muestras de girasol: 22300 (13.27 %)
- Muestras de Maíz (1era cosecha): 10540 (6.27 %)
- Muestras de Maíz (2da cosecha): 4311 (2.56 %)

Entre estas muestras hay múltiples datos por lote adquiridos en distintas fechas, sin embargo no todos los lotes poseen datos en las mismas fechas.

Datos por muestra de entrenamiento	
Nro de lote	Identificador para poder diferenciar lotes entre sí
Fecha	Fecha en la que se adquirieron los datos
Clase	Identificador del tipo de cultivo presente en el lote
Features de banda	Datos estadísticos de las reflexiones obtenidas en cada banda de frecuencia
NDVI	NDVI promedio calculado a partir de las reflexiones obtenidas

Tabla 4.1: Valores presentes en la base de datos para cada muestra

Features por banda (13 bandas por muestra)	
Minimo	Valor mínimo de reflectividad obtenido en el lote
Máximo	Valor máximo de reflectividad obtenido en el lote
Media	Media de los valores de reflectividad obtenidos
Varianza	Varianza de los valores de reflectividad obtenidos
Skewness	Medida de la distorsión de la distribución de los valores de reflectividad respecto a una distribución gaussiana
Kurtosis	Medida que puede indicar presencia de valores atípicos en una distribución

Tabla 4.2: Valores presentes para cada banda en la base de datos.

En primer lugar, se decidió trabajar con los datos de manera independiente al tiempo y lote, es decir, cada muestra se toma como independiente, sin importar que sean datos correspondientes al mismo lote en distinto tiempo. Debido a esta manera de trabajar se decidió categorizar al maíz de primera y segunda cosecha sólo como maíz.

4.2. Métodos no supervisados

Se realizó una prueba de la efectividad de la clasificación utilizando K-Means como algoritmo de clasificación en la base de datos disponible, asignando las clases según la cantidad de datos presentes de cada tipo en los clusters. Esto llevó a un accuracy obtenido fue menor al 40 %. Debido a esto se decidió en adelante centrarse sólamente en los algoritmos clasificados.

4.3. Disminuyendo tiempo de entrenamiento

El objetivo principal de este trabajo fue conseguir los mejores clasificadores supervisados posibles de cultivo. Para esto necesariamente se debe tener la facilidad de poder entrenar los algoritmos en un tiempo razonablemente corto, que permita ajustar los distintos parámetros del modelo, iterar y revisar el impacto del ajuste en la clasificación.

A primera vista se puede observar que la dimensionalidad de los datos no es despreciable, teniendo un total de 79 features por muestra (6 features de banda de cada una de las 13 bandas disponibles, NDVI). El tamaño de la base de datos es un factor a tener en cuenta a la hora de entrenar los modelos. Con esto en mente se consideró la decisión de reducir los datos con los que se trabajará.

Como primera reducción se consideró entrenar los algoritmos sólo con las medias de las reflectancias en cada banda, dado que es una de las medidas que más información sobre una banda pueden dar de las disponibles. Se entrenaron clasificadores con sus parámetros por defecto y se comparó su rendimiento con los mismos clasificadores entrenados sólo con las medias. Este pequeño estudio tuvo como resultados que los clasificadores entrenados con las medias no solo necesitaron un tiempo considerablemente menor para ser entrenados, sino que también obtuvieron un mejor rendimiento.

Con estos resultados se decidió rebajar la dimensionalidad de los datos al tener en cuenta sólo la media de las reflectancias en cada banda. Se analizaron dos posibles maneras de reducir aún más el tiempo de entrenamiento:

- Aplicar análisis de componentes principales (PCA) a los datos.
- Reducir la base de datos, equilibrándola para tener la misma cantidad de muestras de cada clase.

Estas opciones se implementaron y se analizó su efecto en el rendimiento de los clasificadores obtenidos.

4.3.1. PCA

La primera opción para reducir el tiempo de entrenamiento de los algoritmos fue aplicar PCA a la base de datos para reducir la dimensionalidad de las muestras. Se realizó un estudio de la varianza acumulada con cada eje en el caso de aplicar PCA a los datos. Los datos en principio tienen dimensionalidad 14 (La media de cada banda más una feature de NDVI), en la Figura 4.1 se muestran los resultados obtenidos. Se decidió optar por una reducción a 4 dimensiones ya que era suficiente para mantener el 98 % de la varianza de los datos. Luego, se analizaron los resultados obtenidos al entrenar los algoritmos con sus parámetros por default utilizando la base de datos completa y una nueva base de datos reducida con PCA.

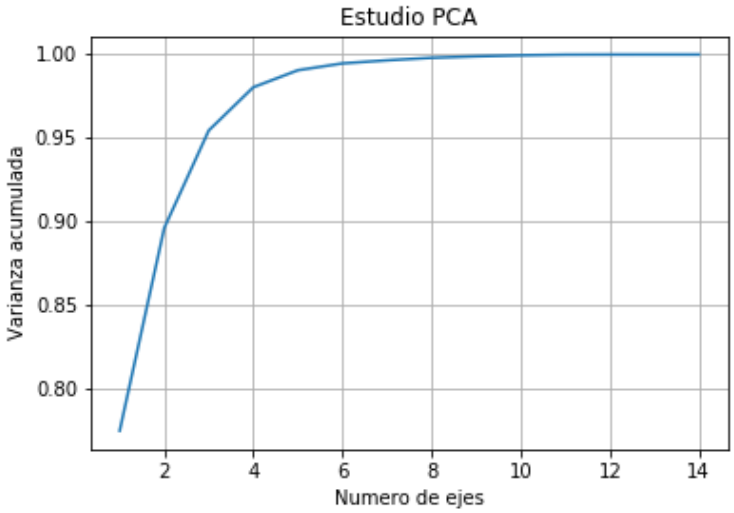

Figura 4.1: Varianza acumulada utilizando PCA en la base de datos.

Las comparaciones se realizaron utilizando las métricas f1-score y kappa, en las Figuras 4.2 y 4.3 se muestran los resultados obtenidos. Una primera observación que se puede dar mirando estos gráficos es que el clasificador SVC consigue un rendimiento muy pobre con los parámetros por defecto. También se puede apreciar que hay una diferencia apreciable en el poder de clasificación de los algoritmos al usar PCA. Si bien la clasificación de soja se mantiene con un f1-score cercano a 0.9 en todos los clasificadores, en el caso de los otros dos tipos de cultivo la métrica empeora drásticamente al usar PCA, salvo en el caso de SVC. Se llegó a la conclusión de que este cambio en el rendimiento se debe al desbalance presente en la base de datos en cuanto a la cantidad de datos de entrenamiento por clase. En la Tabla 4.3 se muestran los resultados obtenidos.

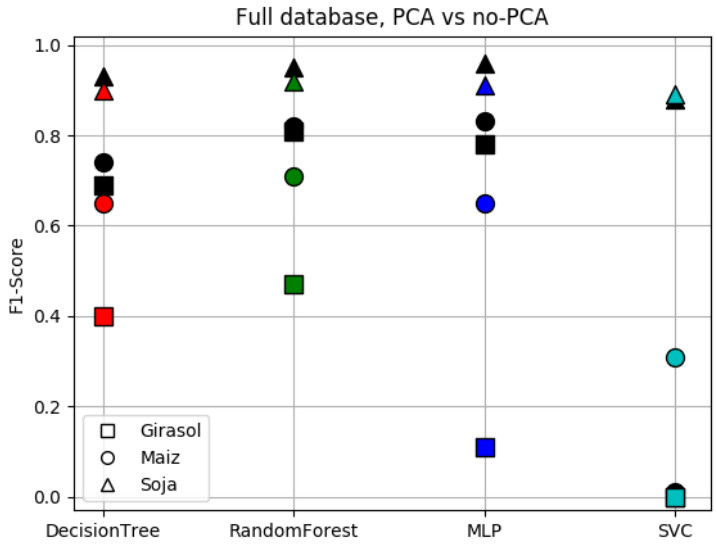

Figura 4.2: F1-score obtenido para cada cultivo utilizando toda la base de datos sin preprocesamiento (marcadores negros) y utilizando PCA (marcadores de colores).

Clasificadores	PCA				NoPCA			
	F1-score			Kappa	F1-score			Kappa
	girasol	maiz	soja		girasol	maiz	soja	
SVC	0.00	0.31	0.89	0.225	0.00	0.01	0.88	0.008
Neural Network	0.11	0.65	0.91	0.538	0.78	0.83	0.96	0.79
Decision Tree	0.40	0.65	0.90	0.540	0.69	0.74	0.93	0.682
Random Forest	0.47	0.71	0.92	0.621	0.81	0.82	0.95	0.784

Tabla 4.3: Valores obtenidos al aplicar y no aplicar PCA a los datos.

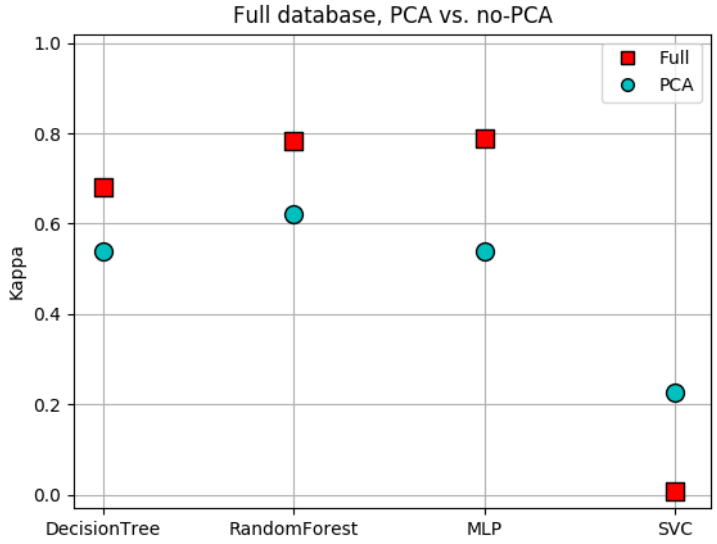

Figura 4.3: Coeficiente kappa utilizando la base de datos completa y aplicando PCA.

4.3.2. Equilibrado de base de datos

La segunda opción para disminuir el tiempo de entrenamiento de los modelos fue disminuir la cantidad de datos presentes en la base de datos, conservando una cantidad comparable de muestras de entrenamiento de cada clase.

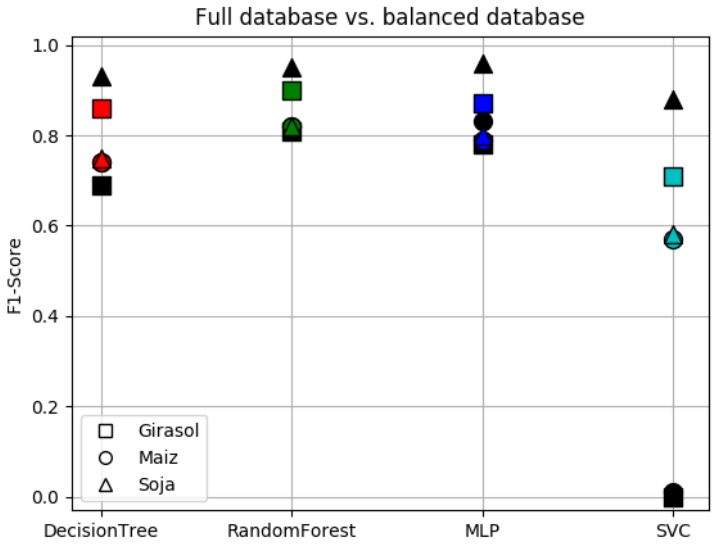

Figura 4.4: F1-score obtenido para cada cultivo utilizando toda la base de datos sin preprocesamiento (marcadores negros) y utilizando la base equilibrada (marcadores de colores).

Se buscó la clase que menos muestras presentaba en la base de datos, resultando ser el girasol con 4311 muestras. Luego se formó una nueva base de datos con 4311 muestras de cada clase y se probaron los modelos en esta base. Los resultados y las comparaciones se muestran en las Figuras 4.4 y 4.5. En estas se puede ver que las métricas no empeoran en gran medida, sin embargo, el tiempo de entrenamiento de los algoritmos se redujo notablemente dado que la cantidad de muestras con la que se entrenan también fue reducida en gran medida.

Los datos completos obtenidos se muestran en la Tabla 4.4.

Clasificadores	Base original				Base equilibrada			
	F1-score			Kappa	F1-score			Kappa
	girasol	maiz	soja		girasol	maiz	soja	
SVC	0.00	0.01	0.88	0.008	0.71	0.57	0.58	0.448
Neural Network	0.78	0.83	0.96	0.79	0.87	0.79	0.80	0.726
Decision Tree	0.69	0.74	0.93	0.682	0.86	0.74	0.75	0.680
Random Forest	0.81	0.82	0.95	0.784	0.90	0.82	0.82	0.774

Tabla 4.4: Valores obtenidos con la base de datos original y la equilibrada.

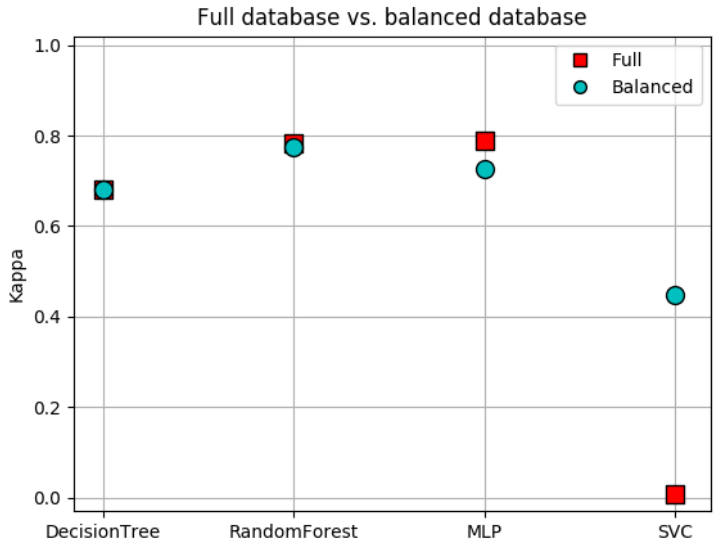

Figura 4.5: Coeficiente kappa obtenido utilizando la base de datos original y la equilibrada.

4.3.3. Conclusiones

Se decidió trabajar con la base de datos equilibrada para buscar los parámetros que lleven a una mejor clasificación de cada algoritmo debido a que con esto se cumple con el requerimiento de que el entrenamiento de los modelos no lleve demasiado tiempo mientras que se no se pierde efectividad de clasificación.

4.4. Barrido de parámetros

Una vez que se eligió trabajar con la base de datos equilibrada se procedió a realizar grillas de parámetros y buscar los que mejor resultados den para cada algoritmo. Esto se realizo mediante Cross-validation de 5 folds, guardando un 20 % de los datos como set de testeo. Se utilizó Kappa como métrica de desempeño durante el cross-validation.

4.4.1. SVC

En el caso del clasificador con vectores de soporte los parámetros que se ajustaron fueron: kernel, C (parámetro de regularización), γ (kernel rbf, inverso del radio de influencia de cada muestra en el entrenamiento) y el grado polinomial (kernel polinomial).

Kernel RBF

Se realizó un barrido de 10 valores equiespaciados logarítmicamente para C y γ entre 1e-4 y 1e4 para C y 1e-2 y 1e6 para γ. Los resultados obtenidos se muestran en

la Figura 4.6. Los mejores resultados se obtuvieron con C=1291 y γ=36. Con estos parámetros se reentrenó el modelo y se lo probó en el test set, obteniendo un accuracy y un weighted f1-score de 91 %, y un kappa de 87.00 %.

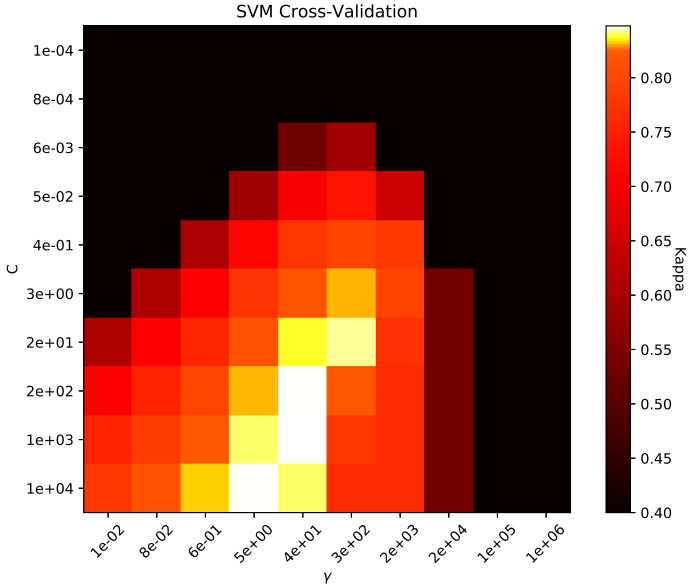

Figura 4.6: Kappa obtenido para la grilla de parámetros explorada utilizando SVM con kernel RBF.

Kernel polinomial

Se realizó un barrido de 10 valores equiespaciados logarítmicamente entre 1 y 1e9 para C, con grados polinomiales entre 1 y 5. Los resultados se muestran en la Figura 4.7. Los mejores parámetros fueron C = 1e9 y grado = 3, con los que se obtuvieron un accuracy de 91 %, Kappa de 86.01 % y weighted f1-score de 91 %.

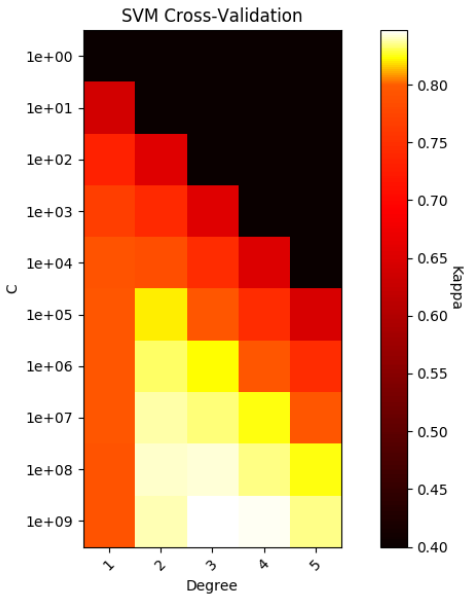

Figura 4.7: Kappa obtenido para la grilla de parámetros explorada utilizando SVM con kernel polinomial.

4.4.2. MLP

En el caso del MLP los parámetros que se ajustaron fueron el parámetro de regularización α, cantidad de capas y neuronas por capa. Se analizaron resultados para una red neuronal de una sola capa y para redes de más capas como casos distintos.

Redes de una capa

En la Figura 4.8 se ven los resultados del barrido realizado. Los mejores parámetros son $\alpha = 0.01$ y 1000 neuronas.

Con estos parámetros se obtuvieron un Kappa de 77.31 %, accuracy y weighted f1-score de 85 %.

Redes de varias capas

En la Figura 4.9 se ven los resultados del barrido realizado. Los mejores parámetros son $\alpha =$ 1e-6 y capas de 200,500 y 1000 neuronas.

Con estos parámetros se obtuvieron un Kappa de 80.57 %, accuracy y weighted f1-score de 87 %.

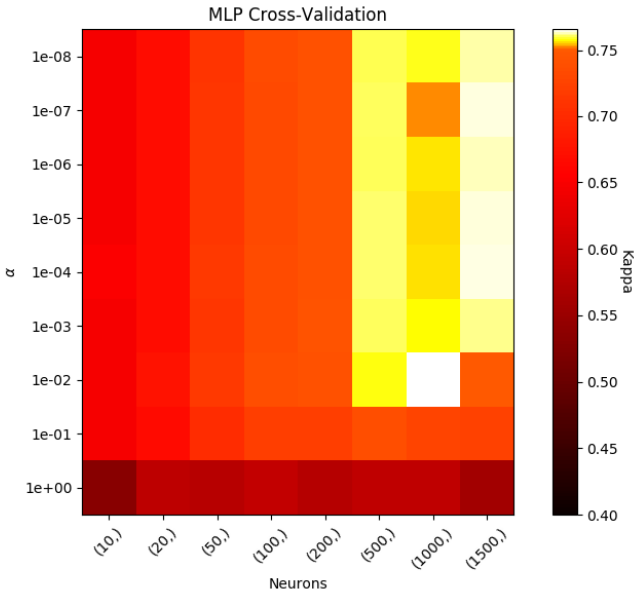

Figura 4.8: Kappa obtenido para la grilla de parámetros explorada utilizando una red neuronal con una sola capa.

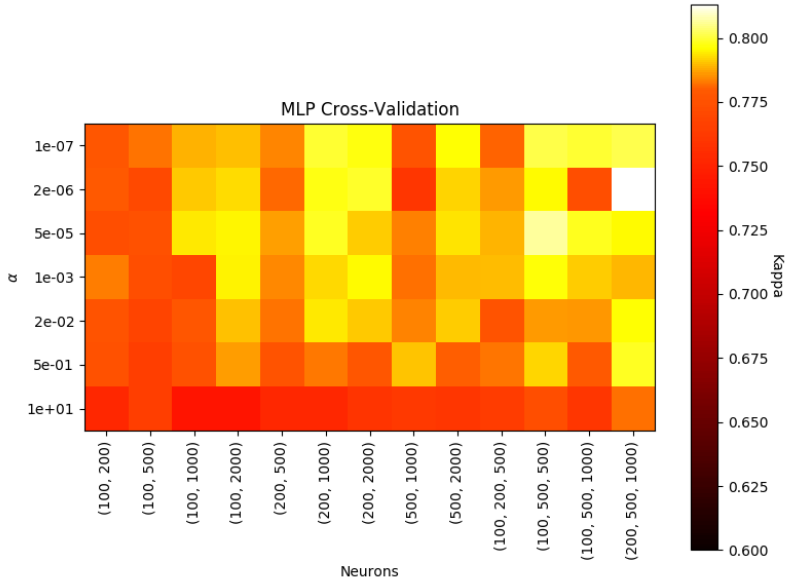

Figura 4.9: Kappa obtenido para la grilla de parámetros explorada utilizando una red neuronal de más de una capa.

4.4.3. Decision Tree

Para el clasificador de árbol de decisión los parámetros que se exploraron son los mostrados en la Tabla 4.5. Con los parámetros que dieron mejores resultados se consiguieron valores de Kappa de 71.17 %, accuracy y weighted f1-score de 81 %.

Parámetro	Valores utilizados
criterion	Gini, **Entropy**
max_depth	5,10,**15**,25,30,None
min_samples_split	2,5,10,**15**,100
min_samples_leaf	1,2,5,**10**

Tabla 4.5: Valores de parámetros utilizados con decision tree. Los mejores valores se encuentran en negrita.

4.4.4. Random Forest

Con este clasificador se utilizaron los mismos valores de parámetros mostrados en la Tabla 4.5, agregando una variación en la cantidad de árboles utilizar. En la Tabla 4.6 se muestran los parámetros junto con los valores con los que se obtuvieron los mejores resultados. Con estos parámetros se obtuvieron un Kappa de 80.80 %, accuracy y weighted f1-score de 87 %.

Parámetro	Valores utilizados
criterion	Gini, **Entropy**
max_depth	5,10,15,25,30,**None**
min_samples_split	**2**,5,10,15,100
min_samples_leaf	**1**,2,5,10
n_estimators	200,**500**,1000,2000,5000

Tabla 4.6: Valores de parámetros utilizados con el clasificador random forest. Los mejores valores se encuentran en negrita.

4.5. Base completa

Una vez que se entendió el funcionamiento de los algoritmos y la dinámica del barrido de parámetros, se probó realizar el mismo barrido que se hizo con el modelo SVM pero utilizando toda la base de datos disponible para analizar los resultados, es decir la base de datos desequilibrada presentada en 4.1. La métrica utilizada para determinar la mejor combinación de parámetros fue el weighted F1-score debido a su aptitud para medir rendimiento en bases desbalanceadas. El mejor rendimiento se obtuvo para los parámetros C=10000 y γ=36, con los que se obtuvo:

- F1-score soja: 96 %
- F1-score maiz: 86 %
- F1-score girasol: 86 %
- Weighted F1-score: 94 %

Se puede observar que el algoritmo priorizó obtener un mejor F1-score para la soja, que es la clase dominante en cuanto a muestras de entrenamiento. Este clasificador puede ser útil para ser aplicado en escenarios donde es esperable encontrar mucha más soja que otros cultivos.

Dado que la base de datos contiene muchos mas datos de soja que de los otros cultivos, se puede pensar que en Argentina es el cultivo predominante. Sin embargo, al buscar información al respecto [27], se encontró que en Argentina la siembra de soja del período 2018/19 ronda las 20 millones de hectáreas sembradas, la de maíz está cerca de las 6.6 millones, y la de girasol las 2 millones. Esto significa que la base de datos no es una muestra representativa de la distribución de cultivos en Argentina, por lo que no sería recomendable utilizarla en su totalidad para entrenar un clasificador que quiera ser aplicado a los cultivos del país. Este problema se puede manejar recortando los datos disponibles o aumentando la cantidad de muestras para que la distribución de cultivos se asemeje a la encontrada en Argentina. Otra forma de manejar el problema es balancear la base de datos.

Se optó por trabajar con la base de datos balanceada porque con esa base se obtendrán clasificadores que no favorecerán a ningún cultivo en específico en la clasificación.

4.6. Conclusiones

Se clasificaron escenas de lotes con cultivos de soja, maíz o girasol. Para esto se analizaron formas de agilizar el entrenamiento de los algoritmos mediante la manipulación de la base de datos y se compararon los resultados obtenidos utilizando la base de datos completa con los obtenidos con una base de datos equilibrada. El clasificador con mejores resultados utilizando la base de datos equilibrada fue el SVC con kernel RBF con los resultados presentados en la Tabla 4.7.

Los resultados obtenidos son mejores que los estudios encontrados en clasificación de cultivos con datos que no tienen en cuenta la variable temporal [15, 16], donde se consiguen clasificadores con kappa menor a 0.85.

El mejor clasificador SVC obtenido utilizando la base de datos sin equilibrar se enfoca principalmente en tener un buen rendimiento en las muestras clasificadas como soja, ya que es la clase que posee la mayor cantidad de muestras. Para aplicar el

	precision	recall	F1-score	Weighted F1-score	Kappa	Accuracy
Girasol	0.94	0.93	0.94	0.91	0.87	0.91
Maiz	0.93	0.87	0.90			
Soja	0.88	0.93	0.90			

Tabla 4.7: Valores obtenidos con el mejor clasificador a nivel objeto sin tener en cuenta la variable temporal.

clasificador en cultivos de Argentina este criterio puede no ser conveniente dado que las proporciones de cultivo de Argentina no son las mismas que las de la base de datos [27].

Capítulo 5

Datos ópticos: Clasificación con datos multitemporales

El siguiente paso en el proyecto fue incluir la evolución temporal de los cultivos en la clasificación. Esto requirió una reestructuración de la base de datos agrupando datos de distintas fechas pero mismo lote en una sola muestra. Para hacer esto se realizó un estudio para determinal con qué nivel de detalle temporal trabajar. Una vez determinado esto se trabajó de la forma ya empleada para conseguir el mejor clasificador posible. A partir de esta parte del trabajo se volvieron a separar las clases de maíz en primera y segunda cosecha.

5.1. Evolución NDVI

En primer lugar se corroboró que exista información disponible en la dimensión temporal. Esto se hizo graficando la evolución temporal del índice NDVI para las distintas clases de cultivo. Se eligió este índice por ser fácilmente interpretable, a diferencia de las reflexiones de una banda en específico.

En las Figuras 5.1, 5.2 y 5.3 se muestran las evoluciones obtenidas para los distintos tipos de cultivo. Se puede apreciar que las evoluciones son diferenciables entre sí salvo en el caso de la soja con el maíz de segunda cosecha. Esto se debe a que los dos cultivos son cosechados en la misma época, por lo que sus curvas de NDVI son parecidas.

Con estos resultados se pudo concluir que efectivamente hay información en la dimensión temporal por lo que puede ser conveniente incluirla en los clasificadores. Tambíen se llegó a la conclusión de que las clases más difíciles de separar parecen ser el maíz de segunda cosecha con la soja.

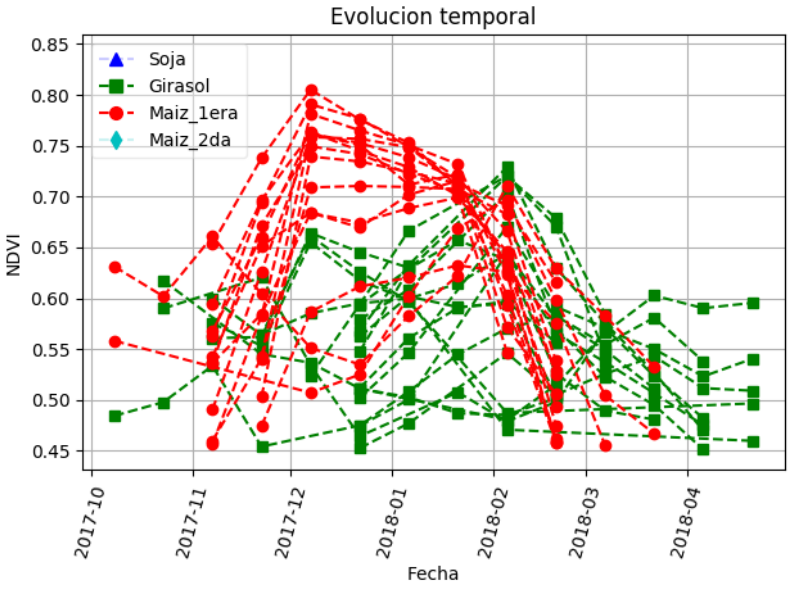

Figura 5.1: Evolución temporal del NDVI en cultivos de girasol y maíz de primera cosecha.

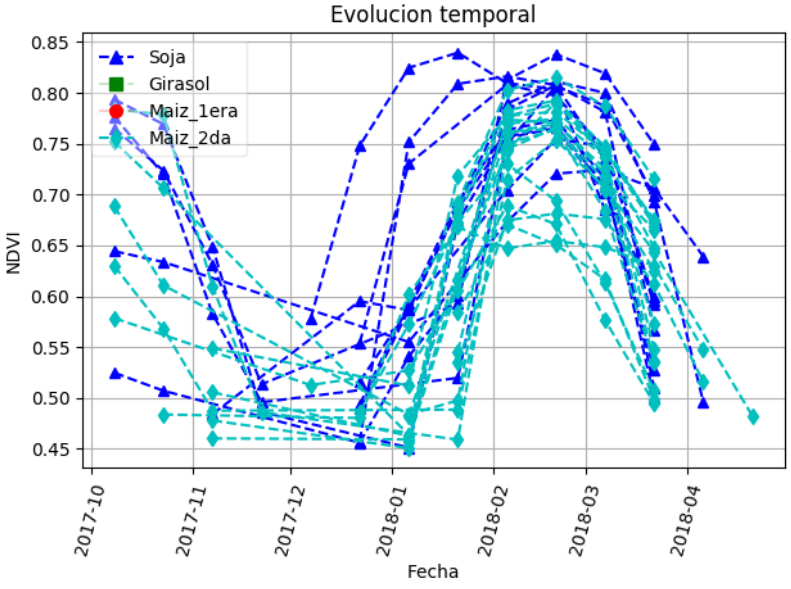

Figura 5.2: Evolución temporal del NDVI en cultivos de soja y maíz de segunda cosecha.

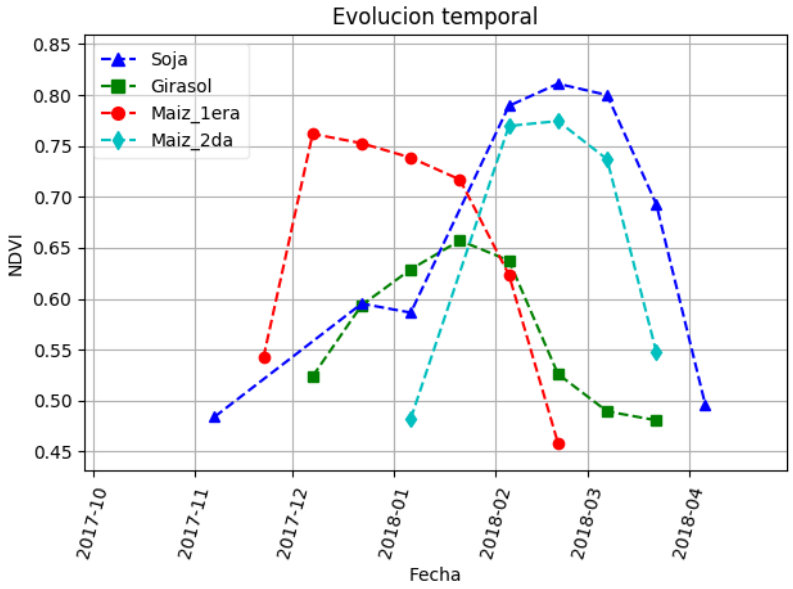

Figura 5.3: Evolución temporal del NDVI en los períodos de crecimiento para todos los cultivos.

5.2. Cambios en la base de datos

Una vez que se decidió incorporar la información temporal en la clasificación fue necesario cambiar el formato de la base de datos para agrupar dentro de una misma muestra los datos obtenidos en distintas fechas para un mismo lote. Los datos disponibles fueron medidos en fechas que van desde octubre del 2018 hasta mayo del 2019. Se realizó la segmentación de este período en grupos de 15 días de duración debido a dos factores:

- La misión Sentinel-2 tiene un tiempo de revisita de aproximadamente 5 días y la disponibilidad de datos depende de condiciones climáticas, por lo que existen muchos días en los que no se dispone de datos.
- La dimensión de las muestras se multiplica por cada muestra temporal considerada, por lo que considerar todos los días implica un aumento en la dimensionalidad innecesario.

La nueva base de datos cuenta con un total de 18827 lotes. Los lotes están distribuidos de la siguiente manera:

- Lotes de soja: 14224 (75.5 % del total)
- Lotes de girasol: 685 (3.6 %)
- Lotes de Maíz (1era cosecha): 3178 (17 %)
- Lotes de Maíz (2da cosecha): 740 (3.9 %)

Nuevamente la base se encuentra claramente desbalanceada, por lo que se trabajó de la misma manera que en el Capítulo 4, balanceándola para obtener una base de datos con 685 muestras de cada clase.

5.3. Barrido de parámetros

Los valores de parámetros por los que se realizaron los barridos fueron los mismos que se usaron en el Capítulo 4. La métrica utilizada para determinar el mejor clasificador fue nuevamente Kappa. Dado que se esperó que la mayor fuente de errores provenga de diferenciar la soja del maíz de segunda debido a sus evoluciones temporales similares (Figura 5.2), se realizaron los barridos utilizando las cuatro clases y eliminando el maíz de segunda para analizar el efecto que tiene en la efectividad.

5.3.1. Utilizando todas las clases

Se realizaron nuevamente los barridos en todos los modelos, los que se reentrenaron con los mejores parámetros observados en estos barridos. En la Tabla 5.1 se muestran los mejores resultados obtenidos con cada modelo. El modelo con los mejores resultados fue el SVC.

	Mejores parámetros	F1-score				Accuracy	Kappa
		G	M1	M2	S		
SVM	Kernel: RBF C: 1291 γ:4.64	0.96	0.97	0.87	0.92	0.93	0.91
MLP	α: 0.01 Capas:(100,2000)	0.93	0.93	0.71	0.82	0.85	0.80
Decision Tree	Criterion: Entropy Max_depth: 10 Min_samples_split: 5 Min_samples_leaf: 1	0.89	0.89	0.71	0.84	0.83	0.78
Random Forest	Criterion: Entropy Max_depth: None Min_samples_split: 2 Min_samples_leaf: 1 N_estimators: 5000	0.95	0.95	0.84	0.90	0.91	0.88

Tabla 5.1: Los mejores parámetros encontrados para cada modelo entrenado con datos ópticos analizado y su rendimiento utilizando todas las clases disponibles.

Observando los valores del F1-score de cada clase se puede concluir que las clases con peor rendimiento son las del maíz de segunda y la soja, como se esperaba, y que la presencia de estas dos clases disminuye el rendimiento de los clasificadores.

5.3.2. Sin maíz de segunda

Se realizaron los barridos sin utilizar los datos pertenecientes a la clase maíz de segunda. Los mejores clasificadores obtenidos junto con su rendimiento se muestran en la Tabla 5.2.

	Mejores parámetros	F1-score			Accuracy	Kappa
		G	M1	S		
SVM	Kernel: RBF C: 1291 γ:4.64	0.96	0.97	0.92	0.93	0.91
MLP	α: 1e-6 Capas:(200,500)	0.97	0.96	0.96	0.96	0.95
Decision Tree	Criterion: Entropy Max_depth: 10 Min_samples_split: 2 Min_samples_leaf: 1	0.92	0.93	0.90	0.92	0.88
Random Forest	Criterion: Entropy Max_depth: None Min_samples_split: 2 Min_samples_leaf: 1 N_estimators: 5000	1.00	0.96	0.97	0.98	0.96

Tabla 5.2: Los mejores parámetros encontrados para cada modelo analizado y su rendimiento sin utilizar los datos del maíz de segunda.

Comparando estos resultados con los de la Tabla 5.1 se puede concluir que el rendimiento obtenido sin utilizar los datos pertenecientes al maíz de segunda mejora en casi todos los modelos.

5.4. Análisis de incorporación de variables

Una vez conseguidos los resultados para cada clasificador se estudió el efecto de cambiar los datos estadísticos de cada banda utilizado. Para esto se entrenaron los modelos con los parámetros con los que se obtuvieron los mejores resultados, pero alterando la información utilizada de la base de datos, y se comparó el kappa obtenido con el obtenido utilizando sólo la media. Los resultados se muestran en la Tabla 5.3. En ella se puede ver que si de desea elegir sólo una medida de las disponibles, la que mejor resultados brinda es la de la media. También se puede concluir que la adición

de features además de la media a lo sumo no empeora el rendimiento, pero no lleva a ninguna mejora.

Features utilizados	SVC	MLP	DT	RF	Mejor	Comparación
Mean	**0.91**	0.80	0.78	0.88	0.91	-
Kurtosis	0.68	0.66	0.55	**0.79**	0.79	-0.12
Skewness	**0.80**	0.77	0.63	0.80	0.80	-0.11
Variance	0.69	0.77	0.67	**0.82**	0.82	-0.09
Mean + Kurtosis	**0.90**	0.79	0.78	0.88	0.90	-0.01
Mean + Skewness	**0.91**	0.79	0.78	0.87	0.91	-
Mean + Variance	0.70	0.79	0.70	**0.84**	0.84	-0.07

Tabla 5.3: Rendimiento obtenido utilizando distintas features y su comparación con el obtenido utilizando sólo la media.

5.5. Conclusiones

Se implementaron clasificadores supervisados para clasificar lotes de cultivo utilizando la información existente en el tiempo. El mejor clasificador conseguido fue un SVC que obtuvo un Kappa de 91 %. Si se compara este clasificador con los resultados de distintos papers [17] [16] [18] [5] [19] en donde sus mejores clasificadores consiguen un kappa que va desde 0.82 a 0.93 se puede decir que el conseguido en este trabajo obtiene un rendimiento comparable con estos, e incluso se encuentra entre los mejores.

También se estudió el efecto de eliminar el maíz de segunda en la clasificación, una clase que posee una evolución temporal similar a la soja y que era la clase con peor rendimiento de todas. Se pudo ver que su eliminación aumentó el rendimiento de los clasificadores, con un clasificador RF obteniendo un kappa de 0.96 y accuracy de 98 %.

Por último se estudió el impacto en el rendimiento de utilizar distintas features, llegando a la conclusión de que es conveniente utilizar sólo la media para la clasificación.

El estudio realizado en este capítulo resalta la eficiencia de tener en cuenta la evolución temporal del cultivo para su clasificación, sin embargo hay que remarcar que para lograr una muestra de este tipo de datos hace falta capturar múltiples imágenes del lote, lo que puede generar inconveniencias. Otro punto a remarcar es que la mayoría de las muestras con las que se trabajó fueron muestras que contienen datos de la época de estudio en su totalidad. Sólo un pequeño porcentaje de las muestras contenían múltiples períodos sin datos.

Si fuese necesario realizar la clasificación de un lote en una fecha previa a su época de cosecha haría falta conseguir datos del lote de días anteriores a dicha fecha. Con estos datos se formaría una muestra donde habrían datos faltantes acorde a la fecha en la que se realice la clasificación. Esta falta de datos puede afectar significativamente

a la eficiencia del resultado de la clasificación. Esta relación entre la falta de datos de una muestra y su eficiencia en la clasificación no se estudió en este trabajo.

Capítulo 6

Agregando datos de radar SAR

Una vez incorporada la información temporal en la clasificación el siguiente paso a realizar fue utilizar datos obtenidos de radar SAR. Se utilizaron datos medidos por la misión Sentinel-1. La cantidad de datos disponibles de este tipo es considerablemente menor a la de datos de imágenes ópticas.

6.1. Datos disponibles

Los datos disponibles de radar para este trabajo consistían en alrededor de 47000 muestras pertenecientes a 3134 lotes distintos, pertenecientes a las provicias de Córdoba y Santa Fe, con clases distribuidas de la siguiente forma:

- Muestras de soja: 34935
- Muestras de maíz de primera cosecha: 9670
- Muestras de maíz de segunda cosecha: 1843
- Muestras de girasol: 525

Nuevamente la base de datos se encuentra desbalanceada a favor de la soja. Se trabajó de la misma manera que en el Capítulo 4 fusionando las dos clases de maíz y equilibrando la base, obteniendo así una base con 525 muestras de cada clase.

En cada muestra se encontraban las features presentadas en la Tabla 6.1.

Features datos SAR	
Polarizaciones	VV, VH
Features por polarización	min, max, media, kurtosis, skewness, varianza
Índices de vegetación	RVI

Tabla 6.1: Features disponibles en la base de datos SAR.

6.2. Datos SAR: Clasificación sin tiempo

Se realizaron los mismos barridos de parámetros y se obtuvieron los resultados mostrados en la Tabla 6.2.

	Mejores parámetros	F1-score			Accuracy	Kappa
		G	M	S		
SVC	Kernel: RBF C: 1e4 γ:35.94	0.71	0.74	0.75	0.73	0.60
MLP	α: 0.01 Capas:(1000,)	0.67	0.65	0.70	0.68	0.52
Decision Tree	Criterion: Entropy Max_depth: 5 Min_samples_split: 10 Min_samples_leaf: 5	0.66	0.72	0.75	0.71	0.57
Random Forest	Criterion: Entropy Max_depth: 5 Min_samples_split: 2 Min_samples_leaf: 1 N_estimators: 200	0.69	0.71	0.74	0.71	0.57

Tabla 6.2: Los mejores parámetros encontrados para cada modelo analizado y su rendimiento utilizando datos SAR y sin tener en cuenta la variable temporal.

Comparando estos resultados con los obtenidos en el Capítulo 4 se puede ver que el rendimiento general es más bajo que con datos ópticos, esto se puede deber a dos factores: La cantidad de muestras en este caso es mucho menor (menos del 15 % de lo disponible en datos ópticos), o que los datos SAR contienen menos información útil para este tipo de problemas. Para poner a prueba la hipótesis de que la cantidad de muestras es importante para el rendimiento se utilizaron 525 muestras de datos ópticos de cada clase para entrenar clasificadores y el rendimiento obtenido de estos se vuelve comparable con los de la Tabla 6.2. Hecho esto se descartó la opción de que los datos de radar contienen menos información que los ópticos y se atribuyó el bajo rendimiento a la baja cantidad de muestras.

6.3. Datos SAR: Clasificación utilizando evolución temporal

Se creó una base de datos similar a la del Capítulo 5 para afrontar el problema de la clasificación con datos multitemporales. La base de datos resultante contenía 2331 muestras de soja, 604 de maíz de primera, 164 de maíz de segunda, y 35 de girasol. La búsqueda de los mejores clasificadores utilizando estos datos obtuvo los resultados

presentados en la Tabla 6.3.

	Mejores parámetros	F1-score				Accuracy	Kappa
		G	M1	M2	S		
SVM	Kernel: RBF C: 21.54 γ: 35.94	1.00	0.88	0.46	0.96	0.93	0.80
MLP	α: 1e-5 Capas:(100,500,500)	1.00	0.89	0.39	0.96	0.93	0.81
Decision Tree	Criterion: Entropy Max_depth: 10 Min_samples_split: 5 Min_samples_leaf: 10	0.86	0.84	0.35	0.95	0.91	0.74
Random Forest	Criterion: Entropy Max_depth: 30 Min_samples_split: 2 Min_samples_leaf: 1 N_estimators: 1000	1.00	0.93	0.45	0.97	0.95	0.86

Tabla 6.3: Los mejores parámetros encontrados para cada modelo analizado entrenado con datos SAR y su rendimiento utilizando todas las clases disponibles.

En esta ocasión el mejor clasificador fue un Random Forest. Todos los clasificadores tuvieron un rendimiento similar al obtenido en el la Sección 5.3.1.

6.3.1. Combinación de datos SAR y ópticos

A continuación se decidió trabajar enfocándose en el problema encontrado en el Capítulo 5 con la diferenciación entre maíz de segunda y soja. Se realizó una comparación de los lotes disponibles de las clases soja y maíz de segunda en las bases de datos multitemporales SAR y de datos ópticos. Se extrajeron los lotes en común de ambas bases, obteniendo dos nuevas bases que contienen datos de 143 lotes para cada clase con la misma agrupación de fechas (15 días por dato) para las distintas fuentes.

Con estas nuevas bases de datos que contienen información de los mismos lotes se entrenaron clasificadores de tres distintas maneras, utilizando datos de Sentinel 1 (S1), utilizando datos de Sentinel 2 (S2) y utilizando ambos tipos de datos, todos teniendo en cuenta la variable temporal. Se buscó el mejor clasificador para cada caso de la misma manera en la que se trabajó en casos anteriores. Los resultados obtenidos con los mejores clasificadores en cada caso se muestran en la Tabla 6.4.

Se pueden apreciar tres cosas importantes:

	S1			S2			S1+S2		
	F1-Score		Kappa	F1-Score		Kappa	F1-Score		Kappa
	S	M2		S	M2		S	M2	
SVC	0.87	0.86	0.72	0.75	0.65	0.41	0.83	0.79	0.62
MLP	0.73	0.71	0.45	0.72	0.61	0.34	0.83	0.79	0.62
DT	0.76	0.72	0.48	0.72	0.65	0.38	0.84	0.81	0.65
RF	0.83	0.79	0.62	0.73	0.64	0.37	0.84	0.81	0.65

Tabla 6.4: Resultados obtenidos con datos multitemporales de distintas fuentes.

- En la mayoría de los casos el rendimiento utilizando los datos de las dos fuentes es mejor que con cualquiera de las fuentes por su cuenta.
- Utilizar sólo los datos SAR llevan a un mejor rendimiento que utilizar sólo los datos ópticos.
- Los rendimientos en general son más bajos que los obtenidos en el Capítulo 5, pero esto es atribuible al bajo número de muestras de entrenamiento.

6.4. Conclusiones

Se realizó la búsqueda del mejor clasificador de cultivo utilizando datos SAR sin tener en cuenta el tiempo y usándolo como una variable más.

El mejor clasificador para los datos de una sola fecha fue un SVC consiguiendo un kappa de 0.60 y un accuracy del 73 %. Este resultado es bajo comparándolo con un estudio que utiliza una red neuronal [20] que consigue clasificar cultivo con accuracy de 86 %, pero comparable con los resultados de otro estudio [16], que consigue clasificar cultivo con accuracy de 75 %. Sin embargo, el clasificador que consigue un 86 % puede haber conseguido este rendimiento debido a las características de los cultivos que se clasificaron.

También se clasificaron muestras con datos donde se usó el tiempo como variable, consiguiendo un accuracy del 95 % y un kappa de 0.86. En este caso se obtienen resultados comparables con los estudios relacionados [16, 20, 21], donde se encuentran clasificadores con un kappa entre 0.68 y 0.87.

Se analizó el impacto del tipo de datos analizado a la hora de diferenciar cultivos con períodos de cultivo similares (maíz de segunda y soja). Se puede concluir que en este caso los datos que más información aportan son los datos de radar, pero que el agregar datos ópticos mejora el poder de discriminación de los clasificadores.

Se pudo concluir que es importante tener una buena cantidad de datos para entrenar un clasificador confiable. También se llegó a la conclusión de que, si existe información de pocos lotes, lo mejor es conseguir datos multitemporales de estos y entrenar un clasificador que tenga en cuenta la evolución temporal de los mismos.

Capítulo 7

Conclusiones generales y discusión

Se pudieron completar los objetivos principales de este trabajo: Estudiar los principios físicos del sensado remoto, aplicar los algoritmos elegidos a la clasificación de cultivo bajo múltiples condiciones y analizar el impacto de las muestras de entrenamiento en la clasificación final.

Respecto a la clasificación de **datos que no tienen en cuenta la variable temporal** el mejor clasificador obtenido utilizando datos ópticos fue un SVC con accuracy del 91 % y un kappa de 0.87. Estos valores son buenos comparados con los estudios encontrados sobre el tema [15] [16]. Para los clasificadores entrenados con datos de radar SAR el mejor clasificador fue un SVC con accuracy del 73 % y un kappa de 0.60. Comparando con uno de los estudios encontrados en el tema [16], se consiguieron resultados comparables, pero comparando estos resultados con los obtenidos en otro estudio [20], se consiguieron bajos rendimientos. Esto último se puede deber a que utilizan una configuración de red neuronal que no se exploró en este trabajo, o a que los cultivos analizados en el estudio son más fáciles de separar que los utilizados aquí. En estas condiciones se puede concluir que la mejor opción para obtener un buen clasificador es utilizar datos de imágenes ópticas.

En el caso de la clasificación de **datos multitemporales** el mejor clasificador entrenado con datos ópticos encontrado fue un SVC con accuracy de 93 % y kappa de 0.91, un rendimiento que se encuentra entre los mejores de los encontrados sobre el tema [17] [16] [18] [5] [19]. Utilizando estos datos se analizó el impacto de eliminar la clase maíz de segunda en la clasificación. Esta clase es díficil de discriminar con la soja para los clasificadores utilizando datos ópticos debido a que poseen épocas de siembra y cosecha similares y los datos ópticos no pueden distinguir características de follaje en los cultivos. Sin esta clase el mejor clasificador fue un RF que alcanzó un accuracy de 98 % y un kappa de 0.96. Esta mejora en el rendimiento de los clasificadores se corresponde con lo esperado. De los clasificadores entrenados con datos SAR el de mejor rendimiento fue un RF con kappa 0.86 y accuracy del 95 %. Este rendimiento se

encuentra entre los mejores comparándolo con los estudios encontrados sobre el tema [16] [20] [21]. Dentro de este caso se analizó cómo afecta el tipo de datos utilizado en cada muestra, dando como resultado que utilizar otras features además de la media no lleva a ninguna mejora en el rendimiento, llegando incluso a bajarlo. Esto se debe a que la sobredimensionalidad de los datos afecta de manera negativa a los algoritmos, y es algo que se aconseja evitar. También se analizó la sustitución de la media por otra medida estadística disponible, lo que dió como resultado que la media es la mejor medidad a utilizar.

La clase con peor rendimiento en ambos casos fue el maíz de segunda así que se decidió hacer una comparación del rendimiento de clasificadores cuya única tarea sea diferenciar el maíz de segunda con la soja. Se entrenaron clasificadores con datos ópticos, SAR y la combinación de ellos, dando como resultado que el rendimiento utilizando datos SAR es mejor que utilizando datos ópticos. Esto era de esperarse dado que al tener fechas similares de siembra y cosecha es crucial poder tener información del follaje para poder diferenciar los cultivos. Por último este estudio dió como resultado que la combinación de los dos tipos de datos es lo que mejor rendimiento obtiene.

Se realizó un análisis de cómo impacta el tamaño, la calidad de los features, y la distribución de las clases en la base de datos, en el rendimiento de los clasificadores. Este análisis llevó a la conclusión de que una base desbalanceada no es recomendable, a menos que este desbalance sea a causa de ser una muestra representativa del contexto en el que se vayan a aplicar los clasificadores, debido a que los algoritmos priorizan la correcta clasificación de las clases que más muestras contengan en la base de datos de entrenamiento. También se llegó a la conclusión de que una sobredimensionalidad de los datos es contraproducente para lograr un buen clasificador.

En general, los clasificadores con mejor rendimiento fueron SVC y RF. Entre los estudios relacionados al tema que se estudiaron el más utilizado es el RF debido a su versatilidad, su robustez ante el overfitting, y la baja complejidad computacional que posee. En general el clasificador SVC obtuvo mejores resultados, pero estos vienen acompañados de alta complejidad computacional a la hora de utilizarlo, por lo que podría ser conveniente utilizar un RF.

7.1. Consideraciones a implementar

Luego de haber completado este proyecto se proponen ideas para realizar si se pretende continuar el trabajo o para tener en cuenta para cualquier trabajo de esta naturaleza que no se pudieron implementar por falta de tiempo:

- Implementar clasificadores jerárquicos [5] con una etapa que clasifique grupos de cultivo por evolución temporal y otra que, definidos los períodos de siembra y

cosecha, discrimine con mayor precisión entre un número reducido de opciones.

- Aumentar el tamaño de la base de datos SAR para poder utilizar una base de datos de gran tamaño que contenga datos combinados y así buscar mejorar el rendimiento de los clasificadores. Esto lleva tiempo debido a que las imágenes deben ser descargadas, preprocesadas y luego ser sometidas al cálculo de features de los lotes presentes en ellas.
- Utilizar imágenes SAR de los satélites SAOCOM, que poseen información en la banda L.
- Implementar otros clasificadores, como redes neuronales profundas o clasificadores bayesianos.

Apéndice A

Actividades de Proyecto y Diseño.

Este trabajo tuvo por objetivo el estudio de técnicas de ciencia de datos en la clasificación de cultivos con imágenes satelitales. Para esto fue preciso entender los principios básicos de adquisición de imágenes satelitales, tanto ópticas como de radar, ya que esto permite abordar la temática de ciencia de datos en forma integrada.

Los algoritmos y programas usados, no son mas que las futuras herramientas que se usaran en todos los procesos de ciencias básicas e ingeniería. Por lo tanto considero que durante todo el trabajo se cumple con la actividad de proyecto y diseños, tanto desde el punto de vista de optimización de recursos computacionales (en este caso crucial) como en el desarrollo de un sistema que clasifica cultivos.

Jorge Lugo.

Bibliografía

[1] Pacala, S., Socolow, R. Stabilization wedges: Solving the climate problem for the next 50 years with current technologies. *Science*, 2004. 1

[2] Peña-Barragán, J. M., López-Granados, F., García-Torres, L., Jurado-Expósito, M., Sánchez de la Orden, M., García-Ferrer, A. Discriminating cropping systems and agro-enviromental measures by remote sensing. *Agronomy for Sustainable Development*, 2008. 1

[3] Peña, J. M., Ngugi, M., Plant, R. Object-based crop identification using multiple vegetation indices, textural features and crop phenology. *Remote Sens. Environ*, 2011. 1

[4] De Wit, A., Clevers, J. Efficiency and accuracy of per-field classification for operational crop mapping. *Int. J. Remote Sens*, 2004. 1

[5] Peña, J. M. e. a. Object-based image classification of summer crops with machine learning methods. *Remote Sensing*, 2014. 1, 15, 27, 46, 53, 54

[6] Blaschke, T. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2009. 2, 27

[7] Orynbaikyzy, A., Gessner, U., Conrad, C. Crop type classification using a combination of optical and radar remote sensing data: a review. *International Journal of Remote Sensing*, 2019. 2, 15

[8] Chuvieco, E. Fundamentals of Satellite Remote Sensing: An Enviromental Approach. 2ª edón. CRC Press, 2016. 2, 5

[9] ESA. Sentinel-1, instrument payload, 2019. URL https://sentinel.esa.int/web/sentinel/missions/sentinel-1/instrument-payload, [Online; accedido 17-09-2019]. 7

[10] ESA. Sentinel-2, 2019. URL https://sentinel.esa.int/web/sentinel/missions/sentinel-2, [Online; accedido 17-09-2019]. 7

[11] Canty, M. J. Image Analysis, Classification and Change Detection in Remote Sensing. CRC Press, 2014. 9, 13, 14

[12] Carletta, J. Assessing agreement on classification tasks: The kappa statistic. *Computational linguistics*, 1996. 11

[13] Arthur, D., Vassilvitskii, S. k-means++:the advantages of careful seeding. *Stanford Semantic School*, 2007. 13

[14] Chan, J., *et al.* An evaluation of ensemble classifiers for mapping natura 2000 heathland in belgium using spaceborne angular hyperspectral (chris/proba) imagery. *Int. J. Appl. Earth Obs. Geoinf.*, 2012. 15, 19

[15] Saini, R., Ghosh, S. Crop classification on single date sentinel-2 imagery using random forest and support vector machine. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018. 19, 38, 53

[16] McNairn, H., Champagne, C., Shang, J., Holmstrom, D., Reichert, G. Integration of optical and synthetic aperture radar (sar) imagery for delivering operational annual crop inventories. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2008. 38, 46, 52, 53, 54

[17] Akhter, S., *et al.* Machine learning approaches on crop pattern recognition, a comparative analysis. *IEEE*, 2018. 46, 53

[18] Song, Q., *et al.* Object-based feature selection for crop classification using multi-temporal high-resolution imagery. *International Journal of Remote Sensing*, 2018. 46, 53

[19] Zhong, L., Hu, L., Zhou, H. Deep learning based multi-temporal crop classification. *Remote Sensing of Environment*, 2019. 46, 53

[20] Del Frate, F., *et al.* Crop classification using multiconfiguration c-band sar data. *IEEE Transactions on Geoscience and remote sensing*, 2003. 52, 53, 54

[21] Xu, L., *et al.* Crop classification based on temporal information using sentinel-1 sar time-series data. *Remote Sensing*, 2018. 15, 52, 54

[22] Bishop, C. M. Pattern Recognition and Machine Learning. Springer, 2006. 15

[23] Peixeiro, M. How to improve a neural network with regularization, 2019. URL https://towardsdatascience.com/how-to-improve-a-neural-network-with-regularization-8a18ecda9fe3, [Online; accedido 04-11-2019]. 18

[24] Scikit-learn. Random Forests, 2019. URL https://scikit-learn.org/stable/modules/ensemble.html#forest, [Online; accedido 04-11-2019]. 19

[25] Sonobe, R., *et al.* Assessing the suitability of data from sentinel-1a and 2a for crop classification. *GIScience remote Sens.*, 2017. 19

[26] USGS. Landsat missions, Landsat 8, 2019. URL https://www.usgs.gov/land-resources/nli/landsat/landsat-8?qt-science_support_page_related_con=0#qt-science_support_page_related_con, [Online; accedido 10-10-2019]. 21

[27] del sur, S. Agricultura argentina, 2019. URL https://surdelsur.com/es/agricultura-argentina/, [Online; accedido 09-10-2019]. 38, 39

Agradecimientos

Gracias a mi familia que me apoyó siempre en las elecciones que tomé respecto a mi vida como estudiante y me seguirá apoyando en la etapa que se viene. A mis compañeros con los que viví innumerables momentos e hicieron de esto una experiencia que valió la pena vivir. Gracias a la gente de INVAP que estuvo siempre para resolver dudas, especialmente Jorge, con quien fuimos aprendiendo de esto a la par.

Agradezco especialmente a mi compañera de vida con quien vivimos muchos altos y bajos, pero siempre al lado del otro. Gracias, Rosita, por ser parte de mi vida.